Universidad Autónoma de Madrid
Escuela Politécnica Superior

Trabajo de Fin de Máster

ALGORITMOS GENÉTICOS PARA LA SELECCIÓN DE VARIABLES EN LA PREDICCIÓN DE PARTIDOS DE BASEBALL

Máster Universitario en Ingeniería Informática

Miguel Vázquez Fernández de Lezeta
Septiembre 2015

ALGORITMOS GENÉTICOS PARA LA SELECCIÓN DE VARIABLES EN LA PREDICCIÓN DE PARTIDOS DE BASEBALL

AUTOR: Miguel Vázquez Fernández de Lezeta
TUTOR: Héctor Menéndez Benito
PONENTE: David Camacho Fernández

Dpto. de Ingeniera Informática
Escuela Politécnica Superior

Universidad Autónoma de Madrid
Septiembre 2015

Resumen

Resumen

La predicción de eventos futuros es una de las tareas más complejas y desafiantes con las que se ha enfrentado el ser humano a lo largo de su historia. Conocer qué va a ocurrir puede ser muy ventajoso para aquellos que poseen la información necesaria. En el caso de los deportes, tener conocimiento de cuán probable es que se dé un resultado concreto en un evento deportivo puede significar un ingreso económico considerable si se sabe aplicar ese conocimiento en el mundo de las apuestas deportivas. Cada vez existen más técnicas de análisis y manejo de datos que permiten predecir eventos futuros, y estas técnicas son aplicadas a las predicciones de resultados de eventos deportivos. Siguiendo la línea de investigación propuesta en el TFG con título Combining Clustering and Time Series for Baseball Forecasting[1], este trabajo busca estudiar técnicas de Minería de Datos aplicadas a predicciones deportivas. Más concretamente, se estudia la aplicación de Algoritmos Genéticos de cara a seleccionar las variables que optimicen el porcentaje de acierto de un Modelo de Predicción, basado en Series Temporales, de partidos de Baseball. El uso de un Algoritmo Genético permite saber qué estadísticas de los equipos, generadas a través de información de distintos partidos, permiten generar Series Temporales que aporten información al proceso de predicción para que esta sea más eficiente.

Palabras Clave

Baseball, Predicción, Minería de Datos, Series Temporales, Algoritmos Genéticos, Estadísticas

Abstract

Predicting future events is one of the most difficult and challenging tasks that human beings have faced along the history. Knowing what is going to happen can provide a big advantage for those who own the proper information. In sports, knowing the probability of a sport event resulting in a concrete outcome may mean a huge benefit if the knowledge is properly applied at gambling. There is a growing number of data analysis and management techniques that allow the prediction of future events, and those techniques are applied to predict the outcome of sports events. Following the research proposal titled Combining Clustering and Time Series for Baseball Forecasting[1], this work aims to study Data Mining techniques applied to sports predictions. More specifically, the application of genetic algorithms in order to select variables to optimize the success rate of a Prediction Model, based on Time Series, for Baseball games. Using a Genetic Algorithm, the system provides information about which team statistics, generated through information from different games, can be used to generate Time Series that provide information to the prediction process in order to increase its efficiency.

Key words

Baseball, Prediction, Data Mining, Time Series, Genetic Algorithms, Statistics

Agradecimientos

Quiero agradecer, especialmente, el apoyo de Carlos, Elena, Natalia y Roberto, porque hemos pasado un mal rato, pero no ha sido solos y lo hemos aguantado con una sonrisa. Quiero agradecer a Héctor, por su ayuda y el haberme seguido guiando y enseñando una vez más, así como a la gente del Departamento de Ingeniería Informática, por el buen ambiente que despliegan y comparten. Por último quiero agradecérselo a mis padres, por aguantar un último año las noches de tecleo intensivo.

Índice general

Índice de figuras

Índice de tablas

1 Introducción

Los sistemas de predicción son unos de los sistemas más complejos del campo del análisis de datos lo cual no es de extrañar, dado que la predicción es una tarea compleja en cualquiera de las áreas de conocimiento en que se aplica, ya sea predicción meteorológica, financiera, deportiva o de cualquier otra clase. El conocimiento aportado por una buena predicción siempre aporta una ventaja a aquel que sabe hacer uso de él, lo que da mucho valor a aquellas herramientas que pueden predecir con certeza y seguriadad qué va a ocurrir en el futuro. En cualquier caso, un proceso tan desafiante como es la predicción no es fácil, ya que muchos factores afectan a la hora de tomar decisiones de cara a las predicciones. Al final, las mejores predicciones son aquellas más acertadas teniendo en cuenta la información disponible y de la cual se parte.

Muchos de los sistemas de predicción conocidos hasta la fecha han sido desarrollados utilizando diferentes técnicas de Minería de Datos como son las Series Temporales[2] y los Algoritmos Genéticos[3], obteniendo resultados notables en campos como la economía o las apuestas deportivas. Existen también muchos estudios que han utilizado previamente Minería de Datos para análisis deportivo[4]. Es importante mencionar que algunos de estos estudios han acabado afectando a los deportes sobre los que estudiaban, por ejemplo traduciéndose a mejoras de los deportistas, a través de la adaptación de nuevos hábitos de entrenamiento o incrementando el rendimiento de los equipos al utilizar estrategias más complejas. Sin embargo, estos estudios no se centran tanto en el campo de la predicción, como en el análisis de deportes para la búsqueda de patrones o estrategias. Esto significa que aún hay un amplio abanico de posibilidades por explorar en cuanto al estudio de técnicas de Minería de Datos aplicadas a la predicción deportiva.

El mundo de las apuestas es una de las áreas más influenciadas por los sistemas de predicción. Conocer los resultados deportivos antes de que se lleven acabo los correspondientes eventos deportivos puede dar un gran beneficio de cara a las apuestas. Las empresas dedicadas a este área necesitan sistemas de predicción robustos para poder establecer precios para las apuestas que puedan resultar atractivos para los clientes, a la vez que se aseguran sacar beneficio del negocio. Cada vez más empresas dedicadas al mundo de las apuestas deportivas, han ido actualizando los sistemas de predicción que utilizan para adaptar técnicas de Minería de Datos, y así poder obtener mayor beneficio del negocio de las apuestas deportivas.

Este trabajo continúa la línea de investigación del trabajo titulado como Combining clustering and time series for baseball forecasting[1], que también dio lugar al artículo Mixed Clustering Methods to Forecast Baseball Trends[5]. En esta línea de investigación se utilizan Modelos de

Predicción de resultados deportivos, más específicamente de baseball, basados en la utilización de técnicas de Minería de Datos. Desarrollando un nuevo Modelo de Predicción basado en el del trabajo anterior, se pretende obtener predicciones más precisas y sólidas. Para ello, se utilizará un Algoritmo Genético en la selección de variables que nos permitan generar Series Temporales, que a su vez se utilizarán para realizar predicciones utilizando el modelo desarrollado. Combinando estas técnicas de Minería de Datos se puede generar un Modelo de Predicción nuevo que pueda ser probado y cuyos resultados puedan ser comparados con los de otros modelos.

El resto del documento se estructura de la siguiente forma: la Sección 2 habla de trabajos similares o relacionados con esta temática, además de explicar algunas de las técnicas de Minería de Datos más relevantes. La Sección 3 describe el nuevo Modelo de Predicción propuesto para este trabajo y cómo se aplica el Algoritmo Genético para la selección de variables a utilizar por el modelo. La Sección 4 muestra todo el proceso de experimentación relacionado a este trabajo, desde la selección de variables como su aplicación utilizando el modelo descrito en la Sección 3. Por último, la Sección 5 muestra las conclusiones del trabajo y da una muestra de los posibles caminos a recorrer en el futuro siguiendo esta línea de investigación.

1.1. Motivación

Las apuestas deportivas, así como el deporte en general, tienen una gran importancia en la sociedad actual a nivel económico. Los grandes eventos y torneos deportivos como los Juegos Olímpicos o Las Grandes Ligas Norteamericanas son seguidos por gente de todo el mundo, ya sean como espectadores, gente involucrada en el mundo de los deportes o incluso corredores de apuestas. El mundo de las apuestas deportivas está extendido a nivel mundial y cuenta con multitud de clientes cuya actividad genera enormes cantidades de dinero para las grandes compañías dedicadas a este tipo de negocios.

El análisis de deportes, así como los estudios estadísticos y los distintos procesos de predicción son interesantes desde el punto de vista de la computación y automatización. Poder predecir de forma certera utilizando grandes conjuntos de datos y en poco tiempo es una tarea compleja, pero interesante, y aún queda mucho por estudiar en este área.

Siguiendo la línea de estudio del trabajo anterior, se quiere utilizar Series Temporales y Algoritmos Genéticos de cara a la predicción para intentar mejorar el Modelo de Predicción existente y así obtener mejores resultados de cara a la predicción, lo cual es desafiante a la par que estimulante.

1.2. Objetivos

El objetivo principal de este proyecto es estudiar cómo un Algoritmo Genético puede aportar información de cara a un Modelo de Predicción de partidos de baseball basado en Series Temporales. Para alcanzar este objetivo se van a llevar a cabo los siguientes procesos:

- **Estudio de técnicas de Minería de Datos** como son las Series Temporales y los Algoritmos Genéticos y cómo se pueden aplicar para extraer información de distintos tipos de datos, específicamente de cara a la predicción.
- **Analizar datos de baseball** y comprender cómo se pueden utilizar de cara al desarrollo de un Modelo de Predicción.
- **Mejora del Modelo de Predicción** analizando el modelo existente, optimizando el código y añadiéndole nuevas características que mejoren el ratio de acierto.

- **Probar y comparar el Modelo de Predicción** y evaluar los resultados obtenidos.

2

Estado del Arte

Esta sección muestra una visión general de la importancia de la industria del deporte, su historia y la importancia económica que tiene. Así mismo, se tratará el uso de metodologías de Minería de Datos y cómo se aplican a los deportes, analizando brevemente herramientas estudiadas, aunque no necesariamente utilizadas en este trabajo.

2.1. Deportes y Sociedad

La industria del deporte es uno de los negocios más rentables del mundo. Los Juegos Olímpicos[6] o la Copa Mundial de la FIFA[7] son ejemplos que muestran la importancia de esta industria en las últimas décadas. Existen otras áreas de estudio involucradas en el mundo de los deportes. Las más importantes son:

1. **Marketing**: La influencia de los deportes en el Mercado es destacable, sobre todo de cara a la publicidad que genera y que aprovechan marcas de diferentes tipos de productos para darse a conocer o ganar prestigio, utilizando la imágen de deportistas y equipos [8].

2. **Medicina**: Muchos avances médicos y estudios del cuerpo humano a nivel físico y fisiológico provienen de la alta inversión de recursos médicos realizada para el tratamiento y la mejora de la salud y eficiencia de deportistas de alto nivel [9].

3. **Tecnología**: De una forma similar a lo que ocurre con la medicina, deportes entre los que destacan el automovilismo o el motociclismo generan inversiones tecnológicas de gran importancia que son aplicadas, no solo al deporte, sino también al desarrollo tecnológico externo a las competiciones deportivas.

4. **Sociología**: Los deportes también tienen influencia en la motivación de países enteros y la psicología social, especialmente cuando un gran evento deportivo de carácter mundial, como los Juegos Olímpicos, es organizado [6].

5. **Economía**: Diferentes deportes o eventos deportivos pueden afectar a la economía de un país entero o a la economía de pequeñas regiones que diseñan productos relacionados a un deporte específico, generalmente relacionado con dicha área [10].

6. **Juego**: Las Loterías Nacionales y las empresas dedicadas a las apuestas también encuentran beneficios de las competiciones deportivas, llegando a influenciar en las decisiones de las Organizaciones Deportivas y en los eventos que llevan a cabo [11].

Además, la predicción de resultados deportivos es uno de los problemas más complejos. El problema consiste en predecir resultados basados en conjuntos de datos extraídos de diferentes deportes, jugadores y partidos de un deporte concreto. Uno de los pasos más complicados de este proceso es encontrar el conjunto de datos apropiado. Normalmente, algunos conjuntos de datos contienen información general de la temporada, mientras que otros aportan información de las jugadas, registros de los partidos, información de las alineaciones, etc. El baseball es uno de los deportes que contiene más información detallada de los diferentes equipos y jugadores. Existe una organización llamada Retrosheet[1] cuya base de datos contiene cantidades ingentes de información sobre jugadores, equipos y partidos, guardando información en registros de los diferentes eventos que ocurren durante un partido.

Existen muchos estudios que se centran en la importancia de los deportes en la sociedad, abarcando los diferentes aspectos sociales en los que influyen, como ya se ha mencionado anteriormente. Algunos trabajos se centran en el efecto de los mayores eventos deportivos a nivel mundial. Por ejemplo, Lee y Taylor [7] analizan la influencia de megaeventos centrándose especialmente en la Copa Mundial de la FIFA de 2002, especialmente en cómo estos eventos crean turismo deportivo, dando un impulso a la economía del país anfitrión (en este caso Corea del Sur). El estudio calcula que la Copa Mundial generó un impacto económico de 1340 millones de dólares en ventas, 307 millones de dólares en ingresos y 731 millones de dólares para el país. Estos resultados muestran cómo el turismo deportivo genera más beneficio que el turismo extranjero de ocio (un 180 % más). Desde un punto de vista distinto, Waitt[6] analiza el impacto social de los Juegos Olímpicos de Sídney, demostrando como la euforia se intensifica desde la organización del evento en 1998 hasta su celebración en el año 2000.

Otros trabajos se centran en el impacto económico de esta industria. Por ejemplo, Pinch y Henry [10] estudian cómo las pequeñas marcas de motores afectan a la economía nacional de Reino Unido. Desde una perspectiva general, Pitts et al. [12] aplica teoría de segmentación de industria a la industria del deporte para crear un modelo para este tipo de industria. En su trabajo utilizan información sobre manufacturación deportiva para hacer una separación en tres categorías: rendimiento deportivo, producción deportiva y promoción deportiva. También estudian cómo clasificar a los distintos tipos de clientes utilizando estas categorías. En este contexto, es importante también estudiar cómo los distintos negocios afectan a los deportes, como se puede ver en el trabajo de Forrest y Simons [11] en el cual se estudia la relación entre deportes y apuestas. En este estudio se centran en cómo las apuestas afectan a las organizaciones deportivas y sus acciones, incluyendo el peligro de la corrupción y cómo estos métodos han afectado a los diferentes deportes, como el cricket, a lo largo de la historia y la influencia de las Loterías Nacionales sobre ellos.

Por último, como en todos los negocios, el deporte necesita de sistemas de medida de calidad para mejorar sus servicios. Ko y Pastore[13] proponen un modelo de calidad de servicio para el deporte, centrando su estudio en cuatro elementos básicos: calidad del programa, calidad de interacción con los clientes, repercusión social y calidad del entorno.

2.2. Equipos y Jugadores

Esta sección analiza la influencia de equipos y jugadores desde diferentes perspectivas, centrándose en diferentes análisis en torno a ambos conceptos individualmente para resaltar

[1] `http://retrosheet.org/`

la importancia que tienen estos factores en el deporte.

2.2.1. Análisis de Equipos

Los equipos pueden ser analizados desde distintas perspectivas. En los negocios deportivos es importante tener en cuenta cómo los equipos afectan a las marcas. Por ejemplo, Gladden y Funk[14] centran su trabajo en entender la dirección de las marcas en los deportes creando un modelo que identifica las diferentes dimensiones de las asociaciones de marcas. En su estudio identifican tres categorías principales: atributos, beneficios y actitud, y su modelo es probado en clientes del mundo deportivo. Desde una perspectiva similar, Bauer et al. [15] estudian la importancia de la imagen de las marcas de cara a la fidelidad de los seguidores en deportes de equipo, mostrando que existen relaciones entre estos factores. De acuerdo con las tres categorías anteriores, la actitud de la marca es lo que más afecta a la fidelidad de los seguidores, lo que se confirma usando un modelado de ecuación estructural.

Para mejorar los equipos existen diferentes estudios acerca de los aspectos relevantes en deportes de equipo. Uno de los más importantes, estudiado por Carron et al. [16], es la cohesión del equipo. En su trabajo tratan de estudiar la relación cohesión y rendimiento en los deportes, a la vez que examinan la influencia de moderadores en este proceso. Con su estudio descubren que la mayor cohesión se da en equipos femeninos. También remarcan la importancia de la cohesión durante la construcción de los equipos y el objetivo de éstos. También es importante la motivación a la hora de mejorar la cohesión, como analizan Roberts y Ommundsen [17] en su estudio sobre cómo la motivación influye en el rendimiento del equipo. Dividen el trabajo en dos metas principales: ego y tareas, estudiando cómo la competición interna afecta al equipo y cómo diferentes factores pueden influenciar en la cohesión de los equipos de acuerdo a estas dos metas. También tratan de determinar los factores de satisfacción de los equipos.

Otro factor importante a tener en cuenta es cómo el entrenador afecta al equipo. Por ejemplo, Gilbert y Trudel [18] estudian la figura del entrenador en equipos deportivos jóvenes. En su trabajo tratan de entender la influencia del entrenador teniendo en cuenta las condiciones del entorno y las características personales, estudiando cómo estudiarlas. De una forma similar, es también importante entender las consecuencias de sustituir a un entrenador. Un ejemplo es el trabajo de McTeer et al. [19], quienes analizan el efecto que causa el cambio de entrenador a media temporada sobre el rendimiento. Su estudio se aplica a cuatro deportes: baloncesto, baseball, hockey y fútbol. Su estudio muestra el efecto del liderazgo de dos formas: cómo un nuevo líder se introduce en el equipo y cómo sus ideas afectan al equipo y también cómo los cambios afectan al rendimiento. En su estudio concluyen que los efectos de este tipo de cambios son mínimos.

Finalmente, se debe entender la importancia de la forma de entrenar al equipo. Baker et al. [20] estudian la influencia de diferentes opiniones de expertos en tres deportes distintos (hockey, netball y baloncesto), tomando información de atletas expertos y no expertos sobre las actividades practicadas durante sus carreras deportivas. Su conclusión es que los expertos dedican 4000 horas a entrenamiento deportivo específico antes de alcanzar un estándar internacional. Otra conclusión que sacan es que existe una correlación negativa entre el número de deportes adicionales y el número de horas dedicadas a entrenamiento específico.

2.2.2. Análisis de Jugadores

Los equipos están formados por jugadores, lo que significa que también es importante estudiar la perspectiva individual del jugador, además de la perspectiva global del equipo.

Uno de los campos de estudio más relevante para los jugadores es la salud. Los deportes requieren de aptitudes físicas extraordinarias y esfuerzos que pueden conducir a lesiones de los jugadores. Un buen ejemplo de la influencia de estas exigencias puede encontrarse en el trabajo de Abdekrim[9], donde se estudian las exigencias físicas producidas por los cambios introducidos en las reglas del baloncesto en Mayo del año 2000, cuando se redujo el tiempo de ataque en 6 segundos y se crearon 4 cuartos en lugar de 2 mitades. En dicho estudio se analiza el efecto sobre el ratio de pulsaciones y la sangre en cuatro posiciones distintas. Generalmente, los jugadores gastan más tiempo en movimientos altamente específicos. Los pivots juegan más intensamente que los aleros y los bases. Los cambios incrementaron ligeramente los efectos cardíacos propios de la competición, y la intensificación difiere de acuerdo a la posición de los jugadores. También es importante estudiar cómo los jugadores se recuperan de las lesiones, como estudian Verral et al. [21]. En su estudio se analiza la pérdida de rendimiento deportivo de los atletas que vuelven al deporte tras la recuperación de lesiones en los isquiotibiales por sobreesfuerzo muscular. Los atletas vuelven en una forma física significativamente más baja inmediatamente tras la lesión, pero el estudio concluye que algunos atletas pueden volver al deporte antes de una completa recuperación de las lesiones, lo cual es necesario para prevenir estos problemas. Otro buen ejemplo en este campo se encuentra en el trabajo de Surve et al.[22], quienes evalúan el efecto de las órtesis semirígidas para tobillos en esguinces de tobillo de jugadores de fútbol. El estudio concluye que las órtesis semirígidas reducen la reincidencia de esguinces de tobillo de jugadores que han sufrido previamente este tipo de lesiones.

Desde un punto de vista social, también es importante estudiar la influencia de jugadores en las marcas de los equipos. Un buen ejemplo se encuentra en el trabajo de Wilson et al.[23], donde se estudia cómo las transgresiones de los jugadores durante el deporte conllevan a repercusiones negativas para las marcas y la gente relacionada a éstas como resultado de su asociación con el deportista. Su estudio se centra en el efecto que estos incidentes pueden tener en relación con patrocinadores desde el punto de vista de una organización deportiva. También discuten diferentes factores acerca del impacto de estas transgresiones.

2.3. Minería de Datos en los Deportes

Las técnicas de Minería de datos se basan en la extracción de conocimiento e identificación de patrones dentro de una fuente de información. Diversas técnicas como estadísticas, aprendizaje automático y Minería de Datos han sido utilizadas para analizar el rendimiento de equipos y jugadores en deportes como el fútbol[24, 25], fútbol americano[26], baloncesto [27, 28, 29], etc. Estas propuestas, generalmente conocidas como modelado de comportamiento humano o robótico, han sido aplicadas en diferentes dominios como simulaciones de fútbol robótico, pero, en estos ejemplos, toda la información es totalmente controlada y simulada. Por otra parte, Raines et al. [30] crean un marco multiagente para analizar comportamientos de equipo. Generan un agente automático para análisis de equipos offline. Esta herramienta tiene diferentes tipos de modelos de comportamiento de equipos para analizar diferentes eventos, como acciones, interacciones y rendimiento. Todas estas metodologías implican modelos de aprendizaje automático enfocados en la generación de interpretación humana. Su dominio de pruebas es el fútbol robótico.

Otro análisis similar aplicado a los deportes de equipo de humanos se encuentra en la liga NBA. Vaz de Melo et al. [29] analizan la evolución de esta liga durante su historia al completo, creando un modelo de redes complejas y estudiando su evolución. En el mismo contexto, Bhandari et al. [27] describen una aplicación de Minería de Datos, llamada Advanced Acout, usada por la National Basketball Association (Asociación Nacional de Baloncesto). Su explicación se centra en cómo esta aplicación ejecuta distintos pasos de la Minería de Datos: extracción

de datos, preprocesamiento de los datos, descubrimiento de patrones y aplicación del modelo. También generan un modelo de visualización basado en patrones y cintas de vídeo. Desde una perspectiva distinta, Ivankovic et al. [31] aplican técnicas de Minería de Datos a información de baloncesto. La meta principal de su trabajo es separar las características más importantes de los datos para predecir resultados de partidos. Así mismo, Biao Xu[32] utiliza un sistema que combina algoritmos genéticos y redes neuronales para predecir el rendimiento deportivo de estudiantes en función de sus atributos físicos y el entorno.

Para el análisis de problemas de fútbol y fútbol americano, Onody y Castro [25] proponen un modelo, también basado en redes complejas pero aplicado únicamente para analizar jugadores Brasileños, y Bitter et al. [28] generan un modelo estadístico, modificando distribuciones probabilísticas clásicas como Bernouilli y la distribución Gaussiana para crear un modelo de puntuación para diferentes ligas. Por otro lado, Dawson et al. [26] estudian los movimientos y actividades de la Liga de Fútbol Australiano usando análisis de vídeo. Extraen información de los patrones de movimiento y actividades de juego estudiando estadísticas de las distintas posiciones. También proponen mejoras en prácticas de entrenamiento específico para diferentes posiciones usando la información extraída.

Finalmente, Shumaker et al. [4] presentan una visión general sobre diferentes aspectos de la Minería de Datos para deportes. Su trabajo se centra en diferentes metodologías, aportando diferentes fuentes de datos de deportes, detalles de búsqueda de diferentes deportes, herramientas para el análisis, modelos de predicción, métodos para analizar contenido multimedia, metodologías para extraer datos de páginas web y algunos datos de estudio usando clasificadores.

2.3.1. Análisis de Baseball

Desde el punto de vista del baseball, hay varios trabajos que se centran en el análisis de baseball. En [33] proponen un marco de visualización de baseball para extraer información sobre diferentes equipos y partidos, de forma que se puedan consultar diferentes aspectos del baseball. Hakes y Sauer [34] estudian el efecto Moneyball desde una perspectiva económica. Su meta es probar que existía una evaluación ineficiente de los jugadores por parte del mercado de fichajes del baseball durante un periodo de tiempo prolongado. La explotación de esta ineficiencia por los Oakland Athletics supuso un progreso destacable para las estrategias del baseball. Otros estudios, como el de Marchi y Albert [35] se centran en diferentes perspectivas analíticas. En este caso introducen varias técnicas para analizar las diferentes partes de un partido de baseball, de equipos, jugadores, etc, usando R. Proveen varios métodos analíticos extraídos de métodos matemáticos. También introducen metodologías de aprendizaje automático pero únicamente centradas en clasificación y regresión.

2.4. Minería de Datos y Herramientas de Análisis de Baseball

Existen muchas herramientas de Minería de Datos que son útiles para el análisis de Baseball. Aquí se enumeran algunas:

- **R**[2]: R es un lenguaje y conjunto de herramientas para computación estadística y gráfica. Bajo la licencia GPL, posee una enorme comunidad que provee diferentes paquetes para análisis estadístico, pudiendo manejar información y big data y ofreciendo reconocimiento de patrones y visualización. Está basado en un lenguaje anterior llamado S y gran parte del código en S funciona inalterado en R.

[2] `http://www.r-project.org/`

- **Matlab**[3]: MATLAB es un lenguaje de alto nivel a la vez que un IDE para computación numérica, visualización y programación. Permite analizar datos, desarrollar algoritmos y crear modelos y aplicaciones. Tiene muchas optimizaciones dedicadas a la computación. Puede ser usado para un gran número de aplicaciones, incluyendo procesamiento de señal y comunicaciones, procesamiento de imagen y video, control de sistemas, medidas y pruebas, computación financiera y computación biológica.

- **Octave**[4]: Octave es un lenguaje interpretado de alto nivel similar a Matlab, principalmente utilizado para computación numérica, bajo la licencia GPL. Se centra en soluciones numéricas de problemas, tanto lineales como no lineales, y experimentos numéricos. También provee una gran capacidad de manejo de gráficos para visualización y manipulación de datos. Este lenguaje es muy similar a Matlab y la mayoría de programas son fácilmente portables.

- **Weka**[5]: Weka es una colección de algoritmos de aprendizaje automático para tareas de Minería de datos. Tiene herramientas para preprocesamiento de datos, clasificación, regresión, clustering, reglas de asociación y visualización. También permite a sus usuarios desarrollar esquemas de aprendizaje automático propios.

- **Improvise**[6]: Improvise es un software basado en arquitectura e interfaz de Java que permite a los usuarios construir visualizaciones interactivamente. Ha sido diseñado para sincronizar diferentes visualizaciones y consultas para combinar información de distintas fuentes. Los usuarios pueden navegar y seleccionar la apariencia de los datos a través de múltiples vistas, usando un número de variaciones en patrones de coordinación bien conocidos como scrolling sincronizado, brushing, drill-down, y zoom semántico.

- **GameDay**[7]: GameDay es un programa software que permite a los seguidores del deporte seguir los partidos de baseball y sus estadísticas en vivo. Para Las Grandes Ligas de Baseball, fue introducido en 2002, un año después de que todas las páginas de los equipos migraran a MLB.com. El software ofrece diferentes herramientas que pueden ayudar en el análisis de partidos.

- **Retrosheet**[8]: Retrosheet es una de las bases de datos más grandes de baseball. También provee diferentes herramientas para analizar los datos que provee. Retrosheen tiene una enorme colección de partidos, en la cual se intenta tener información pública de jugada tras jugada que puede ser interesante para cualquier investigador.

[3] `http://www.mathworks.co.uk/products/matlab/`
[4] `http://www.gnu.org/software/octave/`
[5] `http://www.cs.waikato.ac.nz/ml/weka/`
[6] `http://www.cs.ou.edu/~weaver/improvise/index.html`
[7] `http://mlb.mlb.com/stats`
[8] `http://retrosheet.org/`

3

Modelo de Predicción

Este proyecto se centra en el desarrollo de una metodología de predicción de partidos de baseball basado en series temporales. Para alcanzar este objetivo se utilizarán técnicas de Minería de Datos basadas en series temporales, combinadas con el uso de un algoritmo genético para incrementar el ratio de acierto en las predicciones. Las predicciones se hacen basándose en información recogida de partidos anteriores para que el modelo trate de encontrar similitudes entre partidos y equipos, encontrando así posibles tendencias y patrones que ayuden a predecir el resultado de los partidos. El algoritmo genético se utiliza para seleccionar los valores de los parámetros del modelo de predicción que optimicen el ratio de acierto del modelo. En esta sección se explica cómo funciona el modelo de predicción y cómo se aplica el algoritmo genético en la selección de variables, así como los distintos procesos realizados para llevar a cabo este trabajo.

3.1. Descripción de los Datos

Los datos utilizados para este proyecto han sido proporcionados por Retrosheet[1], organización que se encarga de recoger y actualizar datos de Las Grandes Ligas de Baseball (en inglés Major League Baseball o MLB), como se menciona en la Sección 2. Los datos recogidos por esta organización componen una base de datos de tamaño considerable, donde se recogen datos de partidos desde 1871 hasta la actualidad incluyendo información detallada sobre las jugadas una a una, además de información adicional sobre el baseball americano como detalles de equipos, jugadores, etc.

Dado el carácter cronológico que aporta la información de jugada por jugada que proporciona la base de datos de Retrosheet, así como información de los partidos jugados durante las temporadas profesionales, se pueden obtener estadísticas en función del factor tiempo, lo que permite generar series temporales que se utilizan para el Modelo de Predicción utilizado en este trabajo.

[1] "Los datos utilizados para este proyecto han sido obtenidos de forma gratuita a través de Retrosheet, entidad que posee los derechos de autor de los mismos. Cualquier interesado en estos datos puede contactar con Retrosheet en 20 Sunset Rd., Newark, CP 19711."

3.1.1. Datos de Retrosheet

Los datos sobre partidos de baseball han sido descargados de la página web de Retrosheet. Los ficheros en formato *zip* descargables desde la base de datos de Retrosheet tiene un nombre significativo, compuesto por un número que indica a qué año corresponden los datos y un sufijo indicando qué tipo de datos contienen (para este trabajo se utilizan los ficheros *eve* y *post*, que contienen datos sobre partidos de cada Temporada Regular y los Play-Offs, también conocidos como partidos de postemporada, de cada temporada respectivamente). Por ejemplo, *2003eve.zip* contiene datos de la Temporada Regular jugada durante 2003, mientras que *2003post.zip* contiene información de los partidos de Play-Offs de la temporada de 2003.

Comprimidos dentro de cada fichero en formato *zip* se encuentran tres tipos distintos de ficheros que contienen la información de los partidos de forma separada. Aunque puede encontrarse información más detallada en la página web de Retrosheet[2], a continuación se muestra un resumen con la información contenida en estos ficheros:

- **Ficheros de Evento**: Los ficheros con extensión *.eva*, *.evn* y *.eve* contienen datos de partidos de la Liga Americana, la Liga Nacional y los Play-Offs respectivamente. Los ficheros *.eva* y *.evn* tienen como nombre un año seguido de un código de 3 letras que corresponde a un equipo, y contienen información sobre todos los partidos que dicho equipo ha jugado ese año como equipo local. Para los ficheros con formato *.eve*, sin embargo, en lugar del nombre de los equipos se utiliza una abreviatura del nombre de la fase de los Play-Offs en la cual se juegan los partidos sobre los que se tiene la información. Por ejemplo, *2003OAK.eve* contiene datos de los partidos jugados en 2003 en el campo de los Oakland Athletics, mientras que *2003WS.eve* contiene los datos de los partidos jugados durante la Serie Mundial (World Series) de 2003.

  La información de los partidos contenida en estos ficheros corresponde tanto a características del partido como son los equipos que juegan y sus alineaciones, la fecha y hora en que se juega el partido o las condiciones meteorológicas, además de información detallada de cada jugada como secuencia de bolas, jugadas de bateadores, acciones de corredores y defensores, o sustituciones de jugadores.

- **Ficheros de Equipos**: Estos ficheros no tienen extensión pero son nombrados utilizando el año al que corresponden los partidos y la palabra *TEAM*, y contienen información de cada equipo que juega durante ese año la Temporada Regular o los Play-Offs a los que corresponden los datos (por ejemplo, *2003TEAM*).

- **Ficheros de alineación**: Cada fichero con extensión *.ros* contiene información de la alineación de un equipo durante un año, además de información detallada sobre los jugadores que forman esa alineación, como son posiciones o si son diestros o zurdos de cara al bateo y el lanzamiento. El nombre de estos ficheros se forma con el código de equipo correspondiente y el año. Por ejemplo, *OAK2003.ros* contiene datos de la alineación de los Oakland Athletics durante 2003.

3.1.2. Transformación de los Datos

La base de datos de Retrosheet contiene una enorme cantidad de datos que, sin embargo, deben ser adaptados a un formato con el cual sea viable realizar un procesamiento complejo. Además, de todos los datos que ofrece esta organización, se deben seleccionar los datos que realmente sean relevantes, en este caso aquellos que puedan aportar información a la predicción

[2] `http://retrosheet.org/`

basada en series temporales. Sin embargo, como se trata de una base de datos de gran tamaño, además de que la información está contenida en ficheros con diferentes entradas y campos por entrada, el análisis, extracción y traducción de los datos significativos de cada fichero es una tarea complicada. Sin embargo, se disponen de herramientas software que pueden automatizar y acelerar estos procesos para ahorrar tiempo y recursos, evitando esfuerzos innecesarios. De esta forma la extracción, traducción y limpieza de datos se simplifica enormemente, aunque es importante tener en cuenta que estos procesos requieren de supervisión humana, no solo para identificar los errores que puedan surgir sino para corregir problemas como la pérdida de información o los posibles valores incorrectos que puedan producirse. En este apartado se detalla información de las herramientas que se han usado para llevar a cabo estos procesos automáticos.

Herramientas Software de Retrosheet

Retrosheet ofrece herramientas para trabajar con la información que se puede descargar de su base de datos y transformarla en formatos más sencillos de utilizar en procesado. Para este trabajo se han descargado y utilizado *BGAME.EXE* y *BEVENT.EXE*, que son programas ejecutables para la consola de comandos de cualquier sistema basado en *Windows NT*, y que nos permiten transformar los datos de los ficheros descargables de Retrosheet a ficheros en formato *CSV* que pueden ser importados en una base de datos local. En particular, *BGAME.EXE* genera datos globales de los partidos y *BEVENT.EXE* genera datos de cada jugada. La ventaja de utilizar estas herramientas es que proveen una salida estandarizada de los datos (en lugar de ficheros con información mezclada) y en un formato que otros elementos software son capaces de leer más fácilmente.

Base de Datos MySQL

Como ya se ha mencionado anteriormente, la extracción de datos así como la transformación de los mismos es una tarea compleja que requiere de tiempo y recursos, pero tan importante como estos procesos es la disponibilidad de los datos para su acceso y uso. MySQL es una tecnología de manejo y uso de base de datos muy popular actualmente, que nos permite no solo almacenar los datos que nos interesan localmente sino que además nos proporciona velocidad y versatilidad a la hora de manejar dichos datos. Además de las funciones de manejo de datos propias del lenguaje MySQL, la importación de los datos de los ficheros *CSV* generados por las herramientas software de Retrosheet es una tarea simple, y existen numerosas herramientas y lenguajes de programación que disponen de plug-ins o librerías para conectar con bases de datos MySQL, lo que lo convierte en una herramienta óptima para este trabajo.

3.2. Modelo de Predicción

El Modelo de Predicción utilizado en este trabajo se basa en la comparación de series temporales generadas a partir de las estadísticas deportivas de los distintos equipos de baseball a lo largo de los partidos. Este modelo se basa en la hipótesis de que equipos que se comportan de forma similar a lo largo del tiempo actuarán de una forma similar ante determinadas circunstancias con un alto grado de probabilidad. De esta forma, se estima que los partidos disputados por pares similares de equipos se desarrollarán de una forma parecida.

Una vez los datos han sido extraídos y adaptados para su transformación en series temporales, el modelo es generado. Teniendo información de cada jugada de forma individual, se pueden obtener las estadísticas de los jugadores a lo largo de cada partido y combinarlas para generar las estadísticas de los equipos, considerando estos las sumas de sus jugadores. De esta

forma se obtienen estadísticas a nivel de equipo, tanto jugada por jugada como las estadísticas globales de cada partido a lo largo de las temporadas, que se pueden usar como métricas para la generación de las series temporales que sirven de base para este modelo.

3.2.1. Similitud de equipos

Para este trabajo se han utilizado series temporales basadas en las estadísticas de equipos para obtener el grado de similitud entre equipos. Dicha similitud será la que posteriormente se utilice de cara a la predicción de resultados de los equipos. En primer lugar se crean las series temporales que van a reflejar el rendimiento de los equipos en los partidos que juegan de una forma matemática. Para ello, se han utilizado los partidos que juega un equipo a lo largo de la temporada, ordenados por fecha de juego, como factor cronológico de cara a la generación de las series temporales. Tomando un número P de partidos, se obtienen las estadísticas de los P últimos partidos jugados por cada equipo, generando así series temporales para cada métrica correspondiente a cada equipo a lo largo de toda la temporada, como se muestra en el ejemplo de la Figura 3.1.

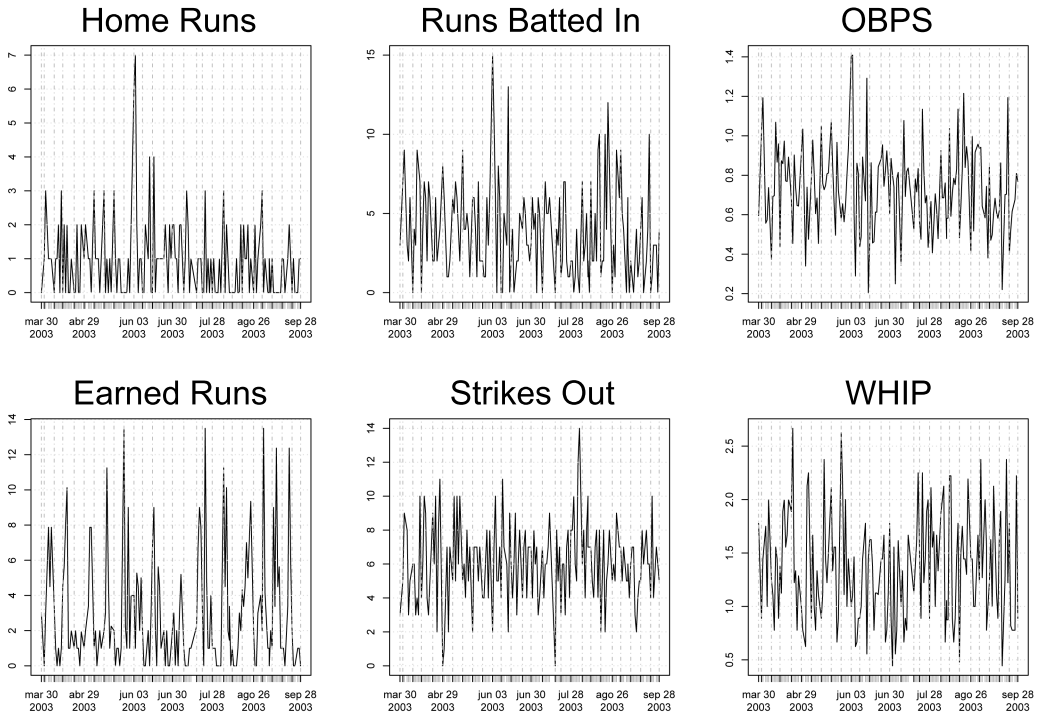


Figura 3.1: Ejemplos de Series Temporales (Anaheim Angels durante la Temporada 2015).

Para obtener la similitud de equipos se realiza un cálculo matemático basado en la disimilitud según la correlación de Pearson estimada de las series temporales par a par[36]. Utilizando la siguiente fórmula se obtiene un valor entre 0 y 1 que nos dice cuán similares son los equipos que estamos comparando:

$$sim(t_i, t_j) = 1 - \left(\frac{\sum_{k=1}^{\Omega} diss(s_k^i, s_k^j)}{Z \cdot \Omega} \right) \tag{3.1}$$

Donde Ω es el número de métricas, t_i, t_j son los equipos que se están comparando, y s_k^i y s_k^j son las series temporales correspondientes de las métricas de los equipos que se están

comparando.Como se explica a continuación, la constante Z es un valor de normalización dado por la función de disimilitud $diss(s_k^i, s_k^j)$ y que corresponde al máximo valor que puede tomar la disimilitud, lo que nos permite normalizar el valor de la similitud a un valor entre 0 y 1. La función de disimilitud se calcula como:

$$diss(s_k^i, s_k^j) = \sqrt{\left(\frac{1-\rho}{1+\rho}\right)^{\beta}} \tag{3.2}$$

Donde ρ es el valor de la correlación de Pearson entre las series temporales s_k^i y s_k^j y β es un parámetro modificable que especifica la regulación de convergencia y nos indica el máximo valor de salida de la función de disimilitud. Dada esta fórmula la función de disimilitud proporciona valores entre 0 y $\sqrt{\beta}$, por lo que si $Z = \sqrt{\beta}$ entonces se obtienen valores entre 0 y 1 para la función de similitud[3].

Simplificación del Modelo

Este modelo corresponde a una simplificación del utilizado en el trabajo previo, Combining clustering and time series for baseball forecasting[1], donde la similitud de equipos se basaba no solo en series temporales, sino en la aplicación de clustering sobre las series temporales para obtener la similitud. El clustering utilizado en este trabajo anterior era un clustering jerárquico basado en la misma disimilitud por correlación de Pearson y se aplicaba sobre agrupaciones de series temporales con la misma métrica pero correspondientes a los distintos equipos, generándose así una jerarquización por métrica. La fórmula que se aplicaba entonces para calcular la similitud era:

$$sim(t_i, t_j) = \frac{\sum_{C_q} \delta_{C_q}^i \cdot \delta_{C_q}^j}{\Omega} \tag{3.3}$$

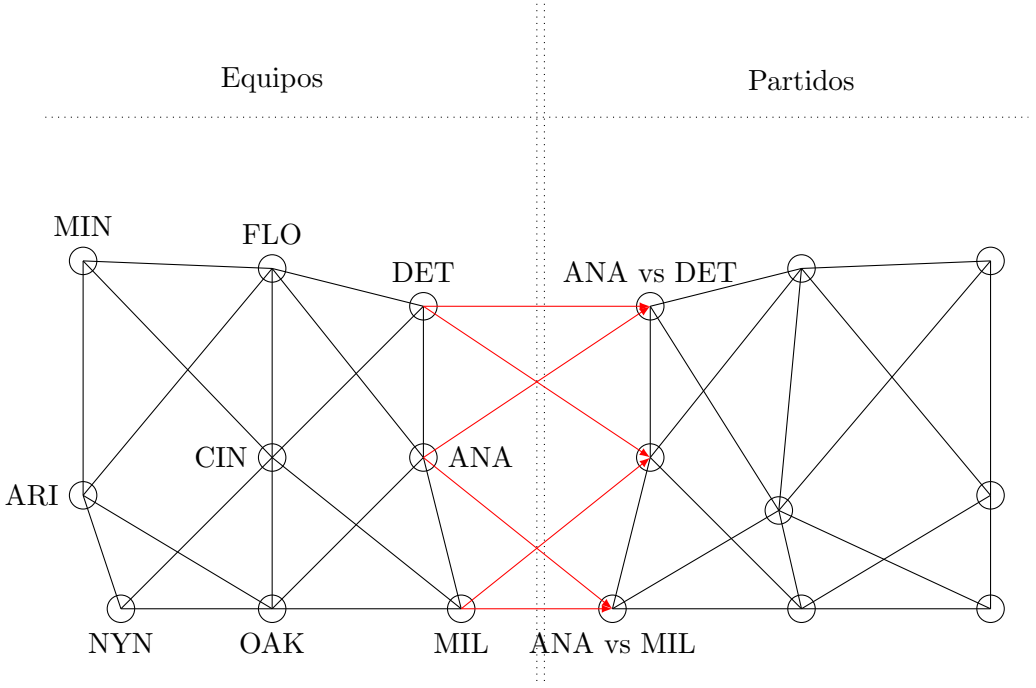
Donde Ω es el número de métricas escogidas, t_i, t_j son los equipos a comparar, C_q representa los posibles clusters por métrica, y $\delta_{C_q}^i$ indica si la métrica se incluye en el cluster o no según la siguiente definición:

$$\delta_{C_q}^i = \begin{cases} 1 & \text{si } t_i \in C_q \\ 0 & \text{en cualquier otro caso} \end{cases} \tag{3.4}$$

La similitud por clustering utilizada anteriormente proporcionaba igualmente un valor de similitud entre 0 y 1. Sin embargo este valor obtenido por comparación de métricas era un valor binario dado, para cada par de equipos, por las series temporales correspondientes a cada métrica indicando únicamente si se encontraban en el mismo cluster o no. Sin embargo, el valor actual provee un valor más preciso de cuan parecidas son estas series temporales, no solo si son parecidas o no, al tratarse de un valor numérico y no un valor booleano. Además, esto supone una limitación ya que la fórmula de similitud por clustering solo puede tener como resultado Ω valores de salida posibles, lo cual hace que sea más difícil determinar qué par de equipos tienen una mayor similitud, sobre todo en casos donde el número de métricas escogidas es pequeño.

Como resultado de esta simplificación, el porcentaje de acierto en la predicción no se ve comprometido y, sin embargo, el tiempo de ejecución de la predicción se reduce considerablemente ya que se ahorra el proceso de clustering de series temporales. Esto es importante dado que el

[3]Dado que en este trabajo se utiliza la función diss.COR[37] del paquete TSclust de R, para este modelo se ha elegido $\beta = 2$ de forma que la disimilitud entre series temporales toma valores entre 0 y $\sqrt{2}$ y por tanto $Z = \sqrt{2}$.



Representación de la similitud entre equipos (izquierda), la similitud entre partidos (derecha) y las conexiones entre equipos y partidos.

clustering se realizaba una vez por métrica en cada predicción y, como veremos posteriormente, durante el experimento se realizan una gran cantidad de predicciones dada la naturaleza del algoritmo genético, lo que supone un ahorro de tiempo global considerable.

3.2.2. Similitud de partidos

Dado que el modelo propuesto tiene como finalidad predecir resultados de partidos de los cuales no se tiene toda la información, es necesario obtener tanta información sobre el partido que se quiera predecir como sea posible, de forma que se facilite el proceso de predicción. Como se ha mencionado anteriormente, este modelo de predicción se basa en la hipótesis de que equipos similares actuarían de forma similar ante circunstancias similares, por lo que se utiliza la similitud de equipos como medio para obtener la similitud entre partidos. Suponiendo partidos distintos, la similitud de éstos vendrá dada por las similitudes entre los equipos que jueguen sendos partidos según la fórmula de similitud basada en disimilitud por correlación de Pearson explicada anteriormente. Así, la similitud entre dos partidos viene dada por la siguiente fórmula:

$$sim(g_i, g_j) = \frac{1}{2} sim(t_h^i, t_h^j) + \frac{1}{2} sim(t_v^i, t_v^j) \tag{3.5}$$

Siendo $sim(t_h^i, t_h^j)$ la similitud de los equipos locales de ambos partidos y $sim(t_v^i, t_v^j)$ la similitud de los equipos visitantes. De esta forma volvemos a tener un valor de similitud en un rango de valores entre 0 y 1 que podemos usar de cara a la predicción, como se explica a continuación.

3.2.3. Predicción de un Nuevo Partido

Para predecir el resultado de un nuevo partido se necesitan datos sobre equipos y partidos que ya hayan sido jugados previamente y cuyos resultados sean conocidos, pues con la información pasada se pueden predecir los comportamientos del futuro. Estos partidos proveen la información necesaria para generar estadísticas y con ello series temporales basadas en el rendimiento de los equipos que han jugado dichos partidos. Para ello, se toma un set de partidos jugados a lo largo de un espacio de tiempo concreto y se generan series temporales en función de las estadísticas de los equipos, generando un conjunto de muestra. Es importante mencionar que para que el modelo funcione, las métricas en que se basan las series temporales extraídas de cada equipo deben ser las mismas para todos, ya que a través de estas se hace la comparación de equipos. Además, cada equipo debe haber jugado como mínimo un número P de partidos en ese periodo de tiempo, pues de otra forma no es posible aplicar la correlación de Pearson, pero este valor P debe ser los suficientemente grande como para que la muestra sea útil aportando información relevante estadísticamente hablando.

Una vez se han creado las series temporales del conjunto de muestra, se generan las series temporales de los equipos que juegan el partido que se quiere predecir. Para ello, se toman los resultados y estadísticas de los últimos P partidos que ha jugado cada uno de los dos equipos antes del partido que se quiere predecir, ya que cuanto más actualizada esté la información, más precisa será la predicción. Utilizando la función de similitud, se haya para cada equipo, tanto para el equipo local t_h como para el equipo visitante t_v, los N equipos del conjunto de muestra que son más similares a ellos. De esta forma se obtiene un conjunto de equipos T_h de equipos similares al equipo local y un conjunto de equipos T_v de equipos similares al equipo visitante.

Teniendo los equipos similares, el siguiente paso es encontrar partidos similares. El modelo propone que para cada par (h_i, v_j), donde $h_i \in T_h$ y $v_j \in T_v$, el partido jugado entre t_h y t_v será similar al partido jugado entre h_i y v_j, y por ello se pueden obtener partidos similares al nuevo partido buscando en el set de muestra los M últimos partidos jugados por cada par de equipos (h_i, v_j). Con este conjunto de partidos extraídos del set de muestra, se puede predecir el resultado del nuevo partido utilizando los resultados de partidos de este conjunto en combinación con el criterio de similitud de partidos en base a la similitud de equipos. La siguiente fórmula calcula la probabilidad de los equipos de ganar el partido en función de dichos criterios:

$$P(t) = \frac{\sum_{i=1}^{G} v(s_i) \cdot sim(g_0, g_i)}{G} \tag{3.6}$$

Donde t es el equipo para el que se quiere obtener su probabilidad de victoria, G es el número de partidos similares extraídos del set de muestra, $sim(g_0, g_i)$ es la similitud entre el partido a predecir y cada partido similar, s_i es cada equipo participante en el correspondiente partido g_i y que es similar al equipo t y $v(s_i)$ es una función definida por:

$$v(s_i) = \begin{cases} 1 & \text{si } s_i \text{ ganó en } g_i \\ 0 & \text{si } s_i \text{ perdió en } g_i \end{cases} \tag{3.7}$$

Finalmente, para decidir qué equipo se debe predecir como ganador, se comparan las probabilidades de ganar de ambos equipos según el cálculo anterior y se predice el equipo con mayor probabilidad como ganador:

$$V(t_h, t_v) = \begin{cases} t_h & \text{si } t_h > t_v \\ t_v & \text{si } t_h < t_v \end{cases} \tag{3.8}$$

La Figura 3.2 muestra un ejemplo esquemático de todo el proceso de predicción. En este ejemplo se seleccionan tres equipos similares para cada equipo que juega el partido a predecir,

donde el color representa cuan similares son los equipos. Con estos dos grupos de equipos y se busca el último partido jugado por cada equipo contra cada uno de los equipos del otro grupo y se obtiene el ganador de cada uno de los partidos. Finalmente, dado que hay más partidos ganados por los equipos similares al equipo local, y que los equipos locales ganadores son más similares que los equipos visitantes ganadores, la probabilidad de ganar del equipo local es mayor y por tanto se predice como ganador.

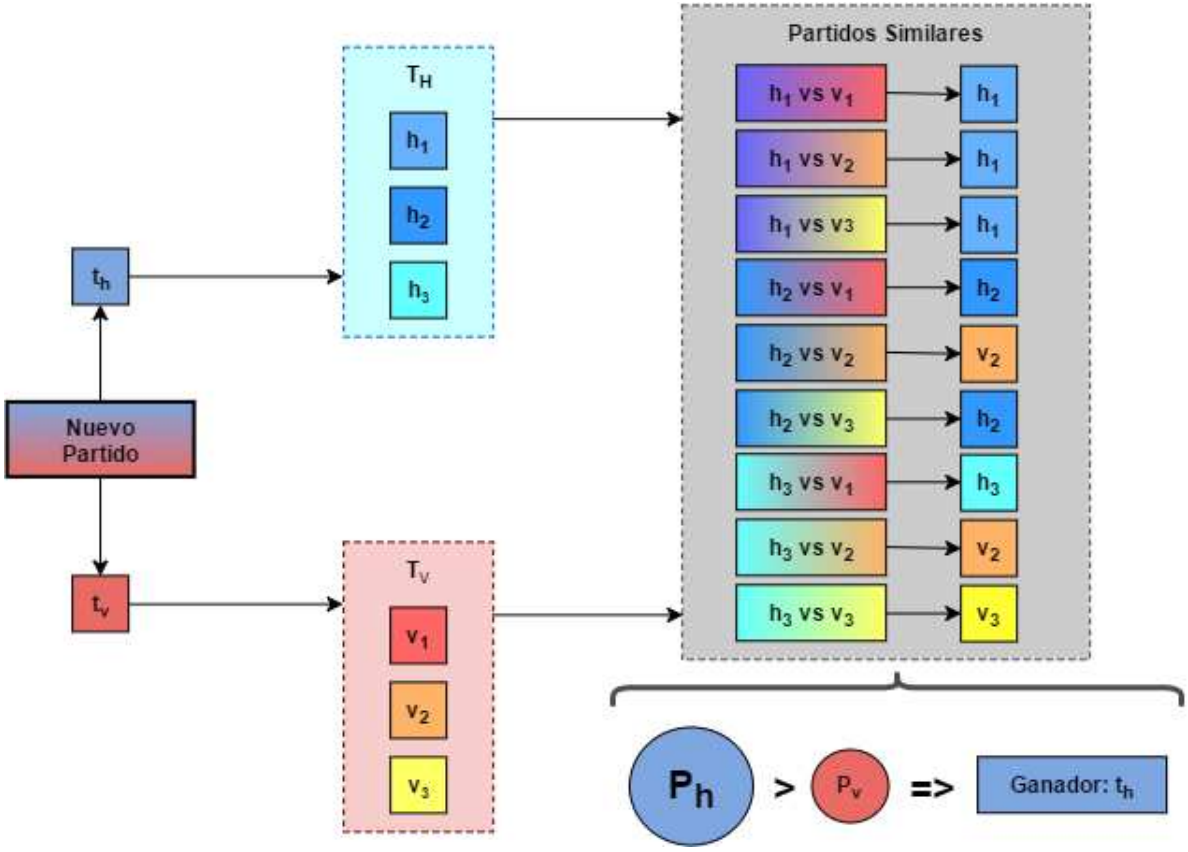


Figura 3.2: Ejemplo esquemático del proceso de predicción.

3.3. Uso del Algoritmo Genético

Una vez desarrollado el modelo de predicción, se necesitan ajustar los parámetros de dicho modelo para la optimización del resultado de las predicciones. Aunque en la Sección 4 se detalla cómo se han elegido las distintas métricas de cara a la experimentación realizada en este trabajo, este apartado se va a centrar en la selección de métricas para la generación de series temporales. En el caso de la fase de experimentación realizada en el trabajo anterior, Combining clustering and time series for baseball forecasting[1], se utilizaban únicamente seis estadísticas como métricas para la generación de series temporales sobre las cuales aplicar el modelo previo para realizar prediciones de partidos. Estas métricas fueron elegidas de entre un conjunto de posibles estadísticas que utiliza la MLB[4] para medir el rendimiento de jugadores y equipos dada su importancia directa de cara al resultado del partido así como su independencia entre ellas.

Sin embargo, existen otras estadísticas publicadas por la MLB que proporcionan información relevante sobre el rendimiento de los jugadores y, aunque haya relaciones matemáticas entre algunas de ellas, pueden aportar más información de cara a optimizar el proceso de predicción.

[4]http://mlb.mlb.com/stats/

Así mismo, algunas métricas pueden estar aportando información redundante o incluso pueden estar afectando negativamente al rendimiento del proceso de predicción, por lo que utilizar todas las métricas simultáneamente no es una buena idea, además de que cuanto mayor es el número de métricas más tiempo requiere el proceso de predicción.

Para seleccionar qué variables optimizan el proceso de predicción de los partidos se va a utilizar un algoritmo genético, de forma que aquellas métricas que proporcionen un mayor porcentaje de acierto sobre conjuntos de partidos a predecir, serán elegidas como métricas para predicciones en experimentaciones posteriores.

3.3.1. El Algoritmo Genético

Los algoritmos genéticos tienen multitud de aplicaciones posibles, por lo que pueden configurarse en un amplio abanico de formas distintas para hacer frente al problema que deban afrontar en cada situación. A continuación se exponen los parámetros más relevantes de un algoritmo genético de cara a su utilización en este trabajo:

- **Tipo de entrada**: El algoritmo soporta distintos tipos de datos sobre los que aplicar modificaciones en busca de la optimización de los mismos. En este caso se utilizan representaciónes binaria de las métricas que se quieren probar para la predicción de los partidos, adjudicando para cada métrica un valor de 0 o 1 en función dependiendo de si la métrica se utiliza en la predicción o no. La ejecución del algoritmo genético hace modificaciones sobre cada representación de las distintas selecciones de métricas, variando el valor asignado a cada métrica y devolviendo finalmente un conjunto de representaciones de las distintas selecciones de métricas y el porcentaje de acierto asignado a dichas selecciones de métricas. Así, las métricas a elegir serán aquellas cuyo valor sea 1 para las representaciones binarias con mayor índice de acierto, mientras que aquellas con valor 0 en las mismas representaciones deberán descartarse.
- **Número de bits**: Para aplicaciones del algoritmo genético sobre entradas de tipo binario, este argumento indica cuantas variables deben ser representadas. En este caso, el número de bits corresponde al número total de métricas que se quieren analizar para su uso en el modelo de predicción.
- **Función de fitness**: Esta función determina cuán bueno es un individuo de la población con respecto al resto. Para este trabajo, la función de fitness corresponde al porcentaje de acierto obtenido tras la predicción de un número A de partidos utilizando el modelo de predicción, lo que significa que a mayor número de aciertos mayor será el valor de la función de fitness y mejor se considerará el individuo. La razón por la cual estos partidos se seleccionan aleatoriamente es para evitar sobreentrenamiento, pues si se seleccionase el mismo grupo de partidos para todos los individuos, el algoritmo acabaría obteniendo el mejor individuo para ese grupo concreto de partidos.
- **Tamaño de la población**: Este valor determina el número de individuos por iteración del algoritmo genético. Teniendo un número P de indviduos por población significa que se disponen de P representaciones binarias de conjuntos de métricas para probar en cada iteración del modelo.
- **Máximo número de iteraciones**: El máximo número de iteraciones determina cuántas iteraciones se realizan en el algoritmo genético como máximo. Esto quiere decir que, a menos que se termine la ejecución por otros métodos, el algoritmo ejecutará tantas iteraciones como se indique en este parámetro (κ), por lo que si se tienen poblaciones de tamaño P se ejecutará la función de fitness un máximo de $\kappa \cdot \mathrm{P}$ veces. Para el modelo

de predicción, dado que la función de fitness ejecuta varias predicciónes (N), el número máximo de predicciones llevadas a cabo será de $\kappa \cdot \mathrm{P} \cdot \mathrm{N}$. Esto es importante de cara a la estimación de tiempos de ejecución del algoritmo, pues el incremento de estos tres parámetros aumenta considerablemente el tiempo de realización de los experimentos.

- **Población inicial**: El algoritmo comienza a ejecutarse sobre una población inicial de individuos sobre los que luego hará las modificaciones oportunas de cara a obtener mejores resultados para la función de fitnes.
- **Función de selección de individuos**: Para cada iteración del algoritmo se necesita una nueva población que es generada utilizando una parte de la población analizada en la iteración anterior. Para ello se hace una selección de los mejores individuos de la población que, como veremos a continuación, aportan información para generar los individuos de la nueva población.
- **Función de cruce**: Una vez seleccionados los individuos a partir de los cuales se va a generar la nueva población de cara a la siguiente iteración, se mezcla e intercambia la información de los individuos seleccionados para generar nuevos individuos.
- **Función de mutación**: Cuando se han generado los cruces de cromosomas, se añade una nueva modificación a los individuos generados para añadir variedad a la población y que los nuevos individuos no dependan completamente de los individuos de la iteración anterior.
- **Elitismo**: El elitismo de un algoritmo genético determina cuántos individuos de una población pasan sin ser modificados a la siguiente población. Esto resulta intersante para mantener al individuo con mejor resultado según la función de fitness dentro de la siguiente población para propiciar que se generen nuevos individuos a partir de este. Como veremos en la Sección 4, para este trabajo se han probado diferentes valores del elitismo dadas las peculiaridades del experimento realizado.

Una vez definido el algoritmo genético con el que trabajar, se puede pasar a la ejecución del mismo para conocer las métricas que optimizan el modelo de predicción. En la próxima sección se describe el proceso de experimentación del algoritmo genético para la selección de variables que se ha realizado para este trabajo. Además, se realizará una segunda experimentación donde las métricas seleccionadas servirán para probar el modelo de predicción propuesto. Es importante mencionar que, aunque este trabajo pretende optimizar el rendimiento del sistema de predicción a través de las métricas óptimas, aún existen muchos parámetros modificables que pueden mejorar el rendimiento del modelo de predicción, como se explica en la Sección 5. De forma similar, los parámetros del algoritmo genético pueden ser también sujetos a modificaciones para optimizar la selección de métricas.

4
Experimentación

Este capítulo muestra cómo se aplican el modelo de predicción y el algoritmo genético sobre un conjunto de datos para analizar la eficiencia de ambos elementos. En esta sección se hablará del conjunto de datos sobre el que se aplicará la experimentación y se expondrán los parámetros escogidos para el modelo explicando las razones por las cuales han sido elegidos.

4.1. Conjuntos de Datos

Para esta experimentación se va a utilizar información de partidos de la base de datos de Retrosheet, como se menciona en la Sección 3. Más concretamente se van a utilizar datos de partidos jugados en la Liga Nacional y la Liga Americana durante la Temporada Regular y los Play-Offs entre los años 2003 y 2011. Las razones por los que estas temporadas han sido elegidas son:

- **Conjuntos de datos de entrenamiento y test**: Para probar el impacto del uso de un algoritmo genético en la búsqueda de variables que optimicen la eficiencia del modelo se han escogido tres conjuntos de datos, que consisten cada uno en datos sobre partidos jugados durante tres temporadas consecutivas, como se muestra en la Tabla 4.1. De esos tres conjuntos, el primero servirá como conjunto de entrenamiento sobre el que aplicar el algoritmo genético y obtener qué variables optimizan la predicción. Una vez obtenidas estas variables, la predicción se aplicará utilizando estas métricas sobre los tres conjuntos de datos, que servirán de conjuntos de test. En este caso se aplica la predicción sobre el propio conjunto de entrenamiento para comparar el porcentaje de acierto con respecto al obtenido para los otros conjuntos y sacar conclusiones sobre posibles sobreentrenamientos.
- **Subconjuntos de datos de muestra**: Como se ha explicado anteriormente, para que el modelo de predicción funcione se utilizan datos de partidos cuyo resultado se conoce para predecir resultados de partidos con resultados no conocidos, obteniendo así un conjunto de muestra y un conjunto de predicción sobre el cual se aplicará el modelo y se comprobará posteriormente el porcentaje de acierto de la predicción respecto al resultado real. Para ello, cada conjunto de tres temporadas mencionado anteriormente se divide en dos partes, dos temporadas que servirán como conjunto de muestra y una temporada que servirá de conjunto de predicción.

- **Impacto del Moneyball**: Se han seleccionado las 9 temporadas siguientes al año 2002, cuando los Oakland Athletics comenzaron a "jugar Moneyball", creando una tendencia a partir de ese momento que se fue adoptando por los distintos equipos de las ligas americanas y los fichajes de jugadores se empezaron a basar en sus estadísticas de juego.

Temporadas	Muestra	Predicción	Uso
2003-2005	2003,2004	2005	Entrenamiento y Test
2006-2008	2006,2007	2008	Test
2009-2011	2009,2010	2011	Test

Tabla 4.1: Conjuntos de datos utilizados para la predicción.

4.2. Selección de variables

En este apartado se detalla el uso de parámetros, tanto para el modelo de predicción como para el algoritmo genético. Para el trabajo anterior, Combining clustering and time series for baseball forecasting[1], se realizaron experimentaciones que ayudan en la selección de algunos parámetros del modelo de predicción, mientras que otros se han elegido siguiendo un razonamiento similar al que se usa en dicho trabajo o adaptándose a las modificaciones y novedades del modelo. Así mismo, el algoritmo genético requiere de parámetros propios para su aplicación.

4.2.1. Selección de Parámetros del Modelo de Predicción

Para que la experimentación sea consecuente, el modelo debe utilizar siempre los mismos parámetros aplicándose sobre distintos conjuntos de datos pues, de otra forma, el entrenamiento del modelo sería incoherente con las pruebas realizadas posteriormente. Como se muestra en la Tabla 4.2, se dispone de varios parámetros que pueden ser modificados, pero la elección de estos debe tener sentido de cara a una experimentación de la cual se puedan extraer conclusiones útiles. A continuación se explica cómo se han seleccionado valores se ha asignado a los distintos parámetros y el razonamiento detrás de estas selecciones.

Las series temporales de este modelo utilizan partidos como la variable temporal y la longitud de éstas debe ser la misma para todas las métricas de cara a poder aplicar la correlación de Pearson. Cada equipo participante en una temporada de Las Grandes Ligas de Baseball juega 162 partidos durante la Temporada Regular, sin embargo hay partidos que se aplazan por diversos motivos, como puede ser la lluvia, y que nunca llegan a retomarse si el resultado de dicho partido no afecta a la clasificación general al final de temporada, haciendo que los equipos implicados hayan jugado solo 161 partidos a final de temporada. Así mismo, hay equipos que juegan más partidos dado que se clasifican para jugar los Play-Offs. Dado que 161 partidos es el mínimo número de partidos que un equipo juega durante una temporada, se ha elegido este valor como longitud de las series temporales, siendo tan grande como para aportar información suficiente de los distintos equipos.

En cada temporada juegan 30 equipos en total (15 en la Liga Americana y 15 en la Liga Nacional), lo que hace que cada conjunto de muestra correspondiente a 2 temporadas disponga de información de 60 equipos, que juegan entre todos unos 4900 partidos por temporada. Durante la predicción, se elige un número de equipos del conjunto de muestra como equipos similares para cada equipo del partido, así como un número de partidos jugado por cada par de equipos similares. Conociendo cómo se realiza el cálculo de probabilidades de victoria de cada equipo, es recomendable que estos valores sean impares, pues reduce la probabilidad de que ambos equipos

obtengan la misma probabilidad de victoria y no se pueda predecir el vencedor. También puede ocurrir, ya que los equipos provienen de dos ligas distintas que no se jueguen partidos entre ninguno de los pares de equipos, en cuyo caso tampoco se puede realizar la predicción. Como se muestra en la Figura 4.1, en la experimentación del trabajo previo se obtuvieron valores para estos parámetros que mejoraban notablemente el porcentaje de acierto de la predicción con respecto a otros (7 equipos y 9 partidos similares), así se utilizan estos mismos valores para esta experimentación.

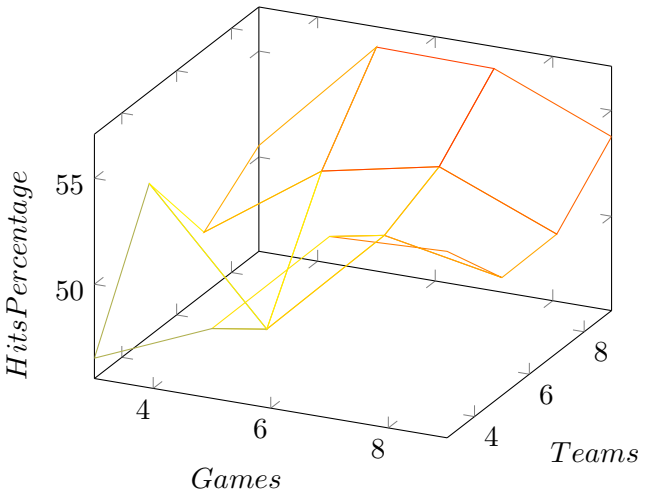


Figura 4.1: Porcentaje de acierto según número de equipos y partidos similares escogidos como parámetro en el modelo original.

Por último, para calcular el valor de la función de fitness del algoritmo genético, para cada individuo, se seleccionarán aleatoriamente 100 partidos del conjunto de predicción y se obtendrá el porcentaje de acierto de las predicciones realizadas con las variables correspondientes a dicho individuo. Dado que, posteriormente, se realizarán experimentaciones con las variables seleccionadas sobre distintos conjuntos de datos, se utilizará este mismo método para calcular la eficiencia del modelo utilizando las métricas seleccionadas por el algoritmo genético.

Parámetro	Valor
Longitud de series temporales (P)	161
Equipos de muestra (E)	60
Partidos de muestra (G)	∼4900
Equipos similares (N)	7
Partidos similares (M)	9
Partidos aleatorios a predecir (A)	100

Tabla 4.2: Selección de parámetros para el modelo.

4.2.2. Selección de Parámetros del Algoritmo Genético

Para este trabajo se utiliza el algoritmo genético disponible en el **paquete GA de R**[38]. Este algoritmo dispone de distintos parámetros a modificar de cara al comportamiento del algoritmo genético, como se expone a continuación.

Para la selección de variables que optimicen el modelo de predicción se ha elegido un conjunto de 40 posibles métricas y en función de ellas se han obtenido las estadísticas de los equipos correspondientes a los partidos de los conjuntos de datos anteriores. Las métricas escogidas se

detallan en el Anexo A. El algoritmo genético selecciona un conjunto de las métricas escogidas como las métricas que optimizan la predicción, y este conjunto será el que se utilice en las experimentaciones consecutivas. Para el algoritmo genético los individuos constituyen secuencias de 40 bit donde cada bit corresponde a cada una de las métricas, como se explica en la Sección 3.3.1.

Para este experimento, se realizarán 100 iteraciones con poblaciones de 50 individuos, lo que supone que la función de fitness se ejecuta 5000 veces, llevándose a cabo un total de 500000 predicciones. Este alto número de predicciones supone un tiempo de ejecución del algoritmo genético notablemente alto, por lo que el código de la ejecución ha tenido que ser revisado y optimizado. Entre otras cosas, la eliminación del clustering del modelo original reduce notablemente el tiempo de ejecución, como se muestra en la Tabla 4.3. Así mismo, el hecho de predecir 100 partidos aleatorios por individuo, en lugar de los cerca de 2430 partidos de la temporada correspondiente al conjunto de predicción, hace que el tiempo de ejecución del algoritmo disminuya considerablemente.

Predicciones	Clustering	Tiempo estimado
100	No	60 segundos
500000	No	3,47 días
100	Sí	110 segundos
500000	Sí	6.37 días
2430	No	24.3 minutos
$12.15 \cdot 10^6$	No	84.38 días

Tabla 4.3: Previsión del tiempo de ejecución según distintos parámetros.

Además de estos valores existen funciones ejecutadas por el algoritmo genético que también son parámetros modificables:

- **Función de fitness**: Como se ha mencionado anteriormente, para calcular el fitness de cada individuo se seleccionan aleatoriamente 100 partidos del conjunto de predicción y se obtiene el ratio de acierto de las predicciones realizadas utilizando las métricas asociadas al individuo.
- **Generación de la población**: Al tratarse de un algoritmo genético aplicado sobre representaciones binarias, la población inicial se compone por secuencias de 40 bits generadas aleatoriamente.
- **Función de selección de individuos**: Para este experimento se utiliza una función de selección que utiliza regresión lineal para elegir qué individuos son seleccionados de cara a la generación de iteraciones posteriores.
- **Función de cruce**: En este caso se utiliza una función de cruce por punto único, lo que significa que, para formar nuevos individuos, se toman dos individuos de la iteración previa, se parten por el mismo punto elegido aleatoriamente, que en este caso, al tratarse de secuencias de bit, se trataría de un punto en dicha secuencia. Una vez partidas ambas secuencias, conocidas como cromosomas, cada parte a un lado del punto de partición de un cromosoma se une a la parte del lado opuesto del otro cromosoma, generando así nuevos cromosomas, como se ve en el ejemplo de la Figura 4.2.
- **Función de mutación**: Para la realización de este experimento se ha escogido un algoritmo de mutación aleatoria que, en función de un parámetro modificable de probabilidad de mutación se modifica el individuo. Para ello, cada bit de la secuencia tiene una probabilidad dada por este parámetro de que su valor pase de 0 a 1 o viceversa.

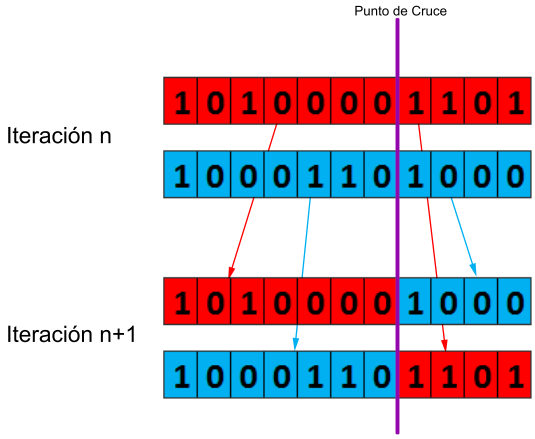


Figura 4.2: Ejemplo de cruce por punto único de cromosomas binarios.

El último parámetro a escoger es el elitismo. En este caso se han realizado dos ejecuciones distintas del algoritmo genético, con elitismo. Esto significa que en la primera ejecución, la población de una iteración es, a priori, completamente distinta de la población de la iteración anterior, mientras que, para un valor de elitismo e, los e mejores individuos pasan a la siguiente población y el resto de la población se completa con nuevos individuos. En este caso el elitismo elegido es 1. Esta doble ejecución se ha realizado para comprobar la evolución de las poblaciones. Al calcularse el fitness de cada individuo en función de 100 partidos aleatorios, en la siguiente iteración los mejores individuos pueden aún mejorar o empeorar su valor de fitness con respecto a iteraciones anteriores. Añadir elitismo consigue que los mejores individuos tengan más oportunidades, lo cual resulta interesante, pero no asegura una mejora del fitness en futuras iteraciones.

Parámetro	Valor
Número de métricas	40
Número de iteraciones	100
Individuos por población	50
Predicciones por individuo	100
Probabilidad de mutación	0.1
Elitismo	0 ó 1
Función de Fitness	Ratio de acierto
Función de Generación	Individuos aleatorios
Función de Cruce	Punto único
Función de Mutación	Probabilidad de bit

Tabla 4.4: Selección de parámetros para el modelo.

4.2.3. Variables Seleccionadas

Una vez ejecutado el algoritmo genético, se ha obtenido un conjunto de individuos con sus porcentajes de acierto asociados. Aunque hay dos ejecuciones distintas, una con elitismo y otra sin elitismo, los resultados no varían mucho, como se puede ver en la Tabla 4.5. La ejecución con elitismo obtiene el mejor resultado, pero, aunque no es mucho mejor que el mejor resultado que la ejecución sin elitismo, mejora notablemente el mejor valor de predicción obtenido con el modelo original (*63.44 %*).

Con los resultados obtenidos en ambos experimentos, se van a escoger varios conjuntos de

Elitismo	Máximo	Media	Desviación Típica
0	0.7222222	0.5381964	0.05102907
1	0.7234043	0.5355111	0.05103607
Ambos	0.7234043	0.5368616	0.05104697

Tabla 4.5: Resultados del algoritmo genético según el elitismo.

métricas siguiendo distintos criterios. Estos conjuntos de métricas serán utilizados para experimentaciones en la segunda parte de la experimentación. Las métricas seleccionadas vienen dadas por:

- Mejores individuos (M1-M10): Se toman las métricas de los 10 mejores individuos por separado.
- Métricas comunes de los mejores individuos (M11-M12): De los 10 mejores individuos se toman aquellas métricas comunes a todos ellos (7 métricas). Así mismo, dado que el número medio de métricas que tienen los mejores individuos es 21, se han elegido las métricas comunes de al menos 6 individuos para tener un número similar de métricas, obteniendo un conjunto de 16 métricas.
- Mejores individuos globales (M13-M22): Dado que algunos individuos aparecen varias veces en el experimento con distintos porcentajes de acierto (por ejemplo, aquellos afectados por el elitismo), se calcula el porcentaje de acierto de cada individuo en el global de la ejecución del algoritmo genético y se toman las métricas de los 10 mejores individuos según este criterio por separado.
- Métricas comunes de los mejores individuos globales (M23-M24): Al igual que en el caso de los mejores individuos, se han tomado las métricas comunes de los 10 mejores individuos globales (1 métrica), así como las comunes a al menos 6 individuos (17 métricas).
- Mejores variables globales (M25-M26): De todos los individuos del experimento, se obtienen aquellas métricas que independientemente tienen mejor resultado. Se han elegido dos conjuntos de métricas, aquellas cuyo porcentaje de acierto está por encima del percentil 75 (10 métricas) y por encima del percentil 50 (20 métricas).

4.3. Resultados del experimento

Una vez obtenidos los distintos conjuntos de métricas según criterios diferentes, se debe comprobar si realmente estas selecciones de variables mejoran la predicción. Para medir la eficiencia de estas métricas, se realizan predicciones de 100 partidos escogidos aleatoriamente, al igual que como ocurría en el algoritmo genético, y se obtiene el porcentaje de acierto. En este caso se va a realizar este proceso 10 veces por conjunto de datos y conjunto de métricas, lo que hace un total de 3000 predicciones por conjunto de métricas. Con una experimentación tan amplia se espera obtener suficiente información como para evaluar correctamente la eficiencia del modelo.

Resultados por Métrica y Temporada

Las tablas 4.6, 4.7, 4.8 y 4.9 muestran los resultados más significativos de la experimentación realizada. En estas tablas se muestran los conjuntos de métricas con los cuales se han obtenido,

Conjunto de métricas	Temporada	Media	Máximo	Mínimo	Desviación Típica
M13	2005	**58.43 %**	70.45 %	48.39 %	±0.065
M1	2005	***57.52 %***	62.11 %	47.31 %	±0.046
M12	2005	56.41 %	61.70 %	48.94 %	±0.049
M14	2008	57.20 %	63.00 %	50.00 %	±0.038
M21	2008	56.50 %	61.00 %	52.00 %	±0.035
M23	2008	56.20 %	62.00 %	51.00 %	±0.035
M7	2011	54.60 %	64.00 %	48.00 %	±0.049
M4	2011	54.10 %	61.00 %	47.00 %	±0.053
M26	2011	53.90 %	64.00 %	46.00 %	±0.048

Tabla 4.6: Métricas con la mejores porcentajes de acierto medios.

para cada conjunto de test, los mejores resultados durante la experimentación. Las tablas con los resultados completos de la experimentación se pueden encontrar en el Anexo B.

La Tabla 4.6 muestra los conjuntos de métricas que han obtenido la mejor media de los porcentajes de acierto para cada conjunto de test, o lo que es lo mismo, el mejor porcentaje de predicción para 1000 partidos aleatorios. Como se puede comprobar, los mejores porcentajes de acierto se encuentran para los conjuntos M13 y M1 aplicados sobre la temporada 2005, además de poseer los mejores resultados máximos de las métricas mostradas. En esta tabla se puede ver que los mejores resultados de los conjuntos de prueba correspondientes a las temporadas 2005 y 2008 no difieren excesivamente, sin embargo para la temporada 2005 las predicciones son mejores y existe una diferencia notable de porcentaje de acierto con respecto a los datos de 2011.

Conjunto de métricas	Temporada	Media	Máximo	Mínimo	Desviación Típica
M13	2005	58.43 %	***70.45 %***	48.39 %	±0.065
M9	2005	53.86 %	65.98 %	43.48 %	±0.067
M3	2005	54.11 %	63.92 %	44.57 %	±0.058
M15	2008	52.90 %	**73.00 %**	37.00 %	±0.098
M22	2008	55.90 %	67.00 %	50.00 %	±0.055
M6	2008	55.70 %	66.00 %	46.00 %	±0.055
M7	2011	54.60 %	64.00 %	48.00 %	±0.049
M26	2011	53.90 %	64.00 %	46.00 %	±0.048
M13	2011	51.00 %	63.00 %	44.00 %	±0.065

Tabla 4.7: Métricas con los mejores porcentajes de acierto máximos.

La Tabla 4.7 muestra los conjuntos de métricas que han obtenido el mejor porcentaje de aciertos para un único conjunto de 100 partidos aleatorios a predecir. Este valor nos indica la máxima eficiencia de cada conjunto de métricas durante el experimento y, de la misma forma, el valor del mínimo porcentaje de acierto nos indica la mínima eficiencia obtenida por el conjunto de métricas. Utilizando este rango de valores se puede comprender variabilidad de la eficiencia de cada métrica.

En la tabla se puede comprobar que los mejores valores se encuentran para los conjuntos de métricas M15, aplicado sobre la temporada 2008, y el conjunto M13, aplicado sobre la temporada 2005. En el caso del conjunto M15, aún obteniendo un valor máximo claramente superior al resto, el valor mínimo asociado es también claramente inferior, lo que significa que el porcentaje de acierto de esta métrica en conjuntos cortos de datos es muy variable, lo cual también confirma un valor de desviación típica alto. Sin embargo, la media de acierto obtenida por dicho conjunto de métricas en esa temporada es baja comparada con los mejores valores

medios. Por contraste, la métrica M13 tiene un valor mínimo no especialmente bajo y, como se veía en la tabla anterior, el valor medio más alto. Así mismo, esta métrica consigue también uno de los valores máximos más altos para la temporada 2011, aunque es notoriamente inferior al obtenido para la temporada 2005.

Conjunto de métricas	Temporada	Media	Máximo	Mínimo	Desviación Típica
M21	2005	55.78 %	62.50 %	***51.69 %***	±0.038
M2	2005	55.23 %	58.95 %	51.06 %	±0.029
M6	2005	55.11 %	61.54 %	50.00 %	±0.038
M21	2008	56.50 %	61.00 %	**52.00 %**	±0.035
M23	2008	56.20 %	62.00 %	51.00 %	±0.035
M8	2008	56.10 %	62.00 %	51.00 %	±0.04
M22	2011	52.30 %	56.00 %	48.00 %	±0.029
M7	2011	54.60 %	64.00 %	48.00 %	±0.049
M8	2011	53.00 %	62.00 %	48.00 %	±0.046

Tabla 4.8: Métricas con los mejores porcentajes de acierto mínimos.

En contrapartida a la Tabla 4.7, la Tabla 4.8 muestra los conjuntos de métricas que han obtenido el mejor mínimo de porcentaje de aciertos para un único conjunto de 100 partidos aleatorios a predecir. Este dato nos permite saber cuánto falla de máximo el conjunto de métricas. En este caso, los mejores valores mínimos han sido obtenidos por el mismo conjunto de métricas M21 para las temporada 2008 y 2011. Esto nos hace pensar que este conjunto de métricas es un conjunto que en caso de tener un ratio de acierto bajo al menos se mantiene en niveles aceptables, y una desviación típica pequeña en ambos casos nos indica que es un conjunto de métricas regular a la hora de ser aplicado por el modelo. Además, en ambos casos la media de acierto es relativamente alta (la media de acierto para la temporada 2008 es la segunda más alta en esa temporada). También se puede mencionar al conjunto M8, que obtiene un valor mínimo considerablemente alto para la temporada 2008 y relativamente alto para la temporada 2011. Además, el porcentaje de acierto de predicción medio es relativamente alto en ambos casos.

Conjunto de métricas	Temporada	Media	Máximo	Mínimo	Desviación Típica
M8	2005	52.20 %	57.89 %	48.91 %	**±0.028**
M2	2005	55.23 %	58.95 %	51.06 %	±0.029
M4	2005	53.18 %	58.33 %	48.31 %	±0.032
M1	2008	52.40 %	59.00 %	48.00 %	±0.034
M23	2008	56.20 %	62.00 %	51.00 %	±0.035
M21	2008	56.50 %	61.00 %	52.00 %	±0.035
M22	2011	52.30 %	56.00 %	48.00 %	***±0.029***
M14	2011	48.80 %	56.00 %	46.00 %	±0.034
M20	2011	47.50 %	54.00 %	42.00 %	±0.039

Tabla 4.9: Métricas con los porcentajes de acierto más estables.

Por último, la Tabla 4.9 muestra aquellos conjuntos de métricas más regulares. La regularidad viene dada por la desviación típica de las 10 iteraciones de 100 predicciones cada una, hechas con cada conjunto de métricas sobre cada conjunto de datos correspondiente a una temporada. De nuevo, el conjunto M8 vuelve a aparecer, pero en este caso para la temporada 2005, donde los ratios de acierto de la predicción medio, máximo y mínimo que presenta son relativamente bajos para esa temporada, pese a la regularidad que ofrece. Lo mismo ocurre con el conjunto M22.

Resultados por Métrica

La Tabla 4.10 muestra los datos de los resultados agrupados por conjunto de métricas. Se dan los mismos datos que en el apartado anterior: media, máximo mínimo y desviación típica de los porcentajes de acierto de las predicciones hechas por el modelo por cada conjunto de métricas distinto. En este caso no se hace separación por temporada y se puede ver el porcentaje de acierto sobre un total de 3000 predicciones por conjunto. Esto es importante puesto que el porcentaje de acierto sobre un conjunto de datos específico (en este caso temporadas) puede no ser significativo de cuán bueno es un sistema de predicción. Por ejemplo, aquellos conjuntos de métricas con buenos porcentajes de acierto para el conjunto de datos referente a la temporada 2005, que se utiliza tanto para entrenamiento como para test, pueden dar información de sobreentrenamiento si no muestran resultados similares al ser aplicados en otros conjuntos de datos.

Conjunto de métricas	Media	Máximo	Mínimo	Desviación Típica
M1	53.61 %	62.11 %	39.00 %	±0.057
M2	53.48 %	65.00 %	45.00 %	±0.049
M3	52.94 %	63.92 %	42.00 %	±0.058
M4	52.96 %	61.00 %	43.00 %	±0.047
M5	50.10 %	58.00 %	40.62 %	***±0.045***
M6	54.37 %	66.00 %	46.00 %	±0.046
M7	**54.54 %**	65.00 %	***47.00 %***	±0.047
M8	53.77 %	62.00 %	**48.00 %**	**±0.041**
M9	51.72 %	65.98 %	41.00 %	±0.063
M10	52.81 %	64.00 %	41.00 %	±0.059
M11	49.34 %	58.00 %	36.00 %	±0.049
M12	54.30 %	62.00 %	44.00 %	±0.051
M13	***54.48 %***	***70.45 %***	44.00 %	±0.065
M14	53.28 %	63.00 %	46.00 %	±0.055
M15	51.72 %	**73.00 %**	37.00 %	±0.073
M16	52.88 %	61.29 %	42.00 %	±0.048
M17	53.45 %	64.00 %	40.00 %	±0.051
M18	53.89 %	63.00 %	45.00 %	±0.047
M19	53.34 %	61.54 %	44.00 %	±0.05
M20	51.76 %	64.00 %	39.00 %	±0.066
M21	53.59 %	62.50 %	43.00 %	±0.052
M22	53.47 %	67.00 %	40.86 %	±0.047
M23	52.59 %	62.00 %	35.48 %	±0.061
M24	52.48 %	60.00 %	38.00 %	±0.046
M25	51.79 %	63.83 %	39.00 %	±0.055
M26	53.79 %	64.00 %	46.00 %	±0.046
Media	*52.94 %*	*63.56 %*	*42.00 %*	*±0.053*

Tabla 4.10: Resultados del experimento por conjuntos de métricas.

Como se puede ver en esta tabla, los mejores valores medios se encuentran para los conjuntos de métricas M7 y M13, donde ambos conjuntos de métricas han mostrado también porcentajes de acierto altos para conjuntos de datos específicos. De una forma similar, el conjunto de métricas M15, que también mostraba un valor medio alto para un conjunto de datos concreto, en esta tabla muestra un porcentaje de acierto cercano al valor medio máximo.

El caso de los valores máximos es obviamente el mismo que para casos individuales, y vemos que M13 y M15 se comportan de una forma similar para conjuntos concretos como globalmente. El caso de valores mínimos cambia relativamente, pues el conjunto de métricas M21 que para

dos conjuntos de datos mostraba mínimos altos, globalmente presenta un mínimo bastante bajo (43 %), lo cual muestra que, en este sentido, no es fiable globalmente. Sin embargo, los mejores valores mínimos vienen dados por los conjuntos de métricas M7 y M8, que ya ofrecían valores mínimos relativamente altos para conjuntos de datos concretos. En cuestión de regularidad, los conjuntos de métricas M8 y M22 vuelven a mostrar desviaciones típicas pequeñas junto con el conjunto M5.

4.3.1. Análisis de los resultados

Con los datos obtenidos tras la experimentación, se puede comparar el modelo de predicción de este trabajo con otros modelos de predicción y comprobar cuán eficiente es con respecto a ellos. Así mismo, se puede analizar si el uso del algoritmo genético para la selección de variables aporta una mejora para que el modelo incremente su ratio de acierto de cara a la predicción de partidos de baseball.

Como se ha explicado anteriormente en esta sección, existen distintos criterios de selección de variables en la utilización del algoritmo genético en un experimento de estas características. Si se toman los porcentajes de predicción obtenidos globalmente utilizando cada conjunto de métricas (porcentaje de acierto de predicción medio), se pueden analizar los conjuntos:

- Mejores individuos (M1-M10): Teniendo en cuenta que estos 10 conjuntos han sido seleccionados como los individuos con mayor acierto a una iteración, podría esperarse que obtuvieran resultados altos y que los primeros conjuntos de este grupo tuvieran resultados mejores que los últimos. Sin embargo, vemos como los resultados, aunque muestren altos porcentajes de acierto en algunos casos (M7 con 54.54 %), en otros son bastante bajos (M5 con 50.10 %) y los distintos valores no están distribuidos uniformemente (M6 y M7 tienen los mejores valores, M1,M2 y M8 valores altos pero no óptimos y M5 y M9 los valores más bajos).

- Métricas comunes de los mejores individuos (M11-M12): Dado que estos conjuntos de métricas se eligen en función de los individuos del grupo anterior, es razonable que se obtengan resultados dispares (49.34 % para 7 variables comunes a todos y 54.30 % para 16 variables comunes a al menos 6 de ellos). Se puede pensar que, en este caso, tener más variables (o un número de variables cercano a 20) aporta más información útil a la predicción.

- Mejores individuos globales (M13-M22): En este caso y a diferencia del grupo de mejores individuos según una iteración única, los individuos de este grupo tienen un porcentaje de acierto similar entre ellos y relativamente alto (7 de 10 por encima del 53 %). Además, el conjunto M13 que se presuponía mejor según la experimentación con el algoritmo genético, demuestra muy buenos resultados, tanto en valor promedio como en valor máximo.

- Métricas comunes de los mejores individuos globales (M23-M24): La selección de métricas utilizando aquellas comunes para los individuos del grupo anterior en este caso muestran, al igual que entre los individuos, un comportamiento similar independientemente del número de métricas utilizadas. Sin embargo, el resultado baja del 53 % en los dos casos, lo que puede significar que, para este modelo, las métricas no afectan independientemente al proceso de predicción.

- Mejores variables globales (M25-M26): Con los resultados aportados estos dos conjuntos de métricas se puede concluir, al igual que con los conjuntos M23 y M24, las variables aportan información a la predicción por dependencias entre ellas, pues el mejor resultado de las variables escogidas independientemente no alcanza el mejor resultado obtenido por

un individuo concreto. También, al igual que con los conjuntos M11 y M12, se puede pensar que 20 métricas obtienen mejores resultados que 10 porque aportan más información sin ser excesiva.

Tras este análisis, se van a elegir los conjuntos M7 y M13 como los conjuntos de métricas que debe utilizar el modelo de predicción, ya que tras el experimento ambos muestran los mayores porcentajes de acierto global en la predicción. Si comparamos el modelo de predicción utilizando estos conjuntos de métricas con respecto a otros modelos de predicción se pueden sacar conclusiones interesantes.

Por ejemplo, el modelo de predicción del trabajo anterior, Combining clustering and time series for baseball forecasting[1], para el conjunto de datos correspondiente a de partidos de las temporadas jugadas entre 2003 y 2005 y utilizando parámetros similares obtenía un porcentaje de acierto del 56.04 % de media y 62.64 % máximo, mientas que este nuevo modelo utilizando el conjunto de métricas M17 obtiene un 54.52 % de media y un 62.77 % máximo y con el conjunto M13 obtiene un 58.43 % de media y un 70.45 % máximo. Podemos pensar que el modelo M13 mejora el modelo, pero podría tratarse de sobreentremaniemto.

Suponiendo un modelo que predijese siempre al equipo local como vencedor, se obtienen porcentajes de acierto que podemos comparar con los obtenidos por nuestro modelo, como se muestra en la Tabla 4.11. Esta tabla que ambos conjuntos mejoran el porcentaje de acierto sobre el global, pero se ve como el conjunto M13 mejora mucho el modelo para los datos de la temporada 2005, pero no mejora para los otros dos conjuntos de datos, lo que indica la existencia de sobreentrenamiento. Sin embargo, el conjunto de métricas M7 muestra un ratio de acierto constante que supera, en la mayoría de los casos, el ratio del modelo de predicción que da siempre al equipo local como vencedor.

Temporada	Victoria Local	Modelo con M7	Modelo con M13
2005	53.75 %	54.52 %	58.43 %
2008	55.65 %	54.50 %	54.00 %
2011	52.65 %	54.60 %	51.00 %
Total	54.02 %	54.54 %	54.48 %

Tabla 4.11: Comparación del Modelo de Predicción utilizando el conjunto de métricas M13 y un Modelo de predicción por prioris.

4.4. Discusión de los Resultados

Los resultados de este experimento muestran cómo el uso de un algoritmo genético puede ayudar en la selección de variables para modificar el comportamiento de un modelo de predicción. Aunque el modelo pueda llegar a mejorar, se debe tener cuidado para no caer en prácticas que sobreentrenen el modelo, haciéndolo menos efectivo.

El modelo planteado representa para distintos conjuntos de variables comportamientos dispares, aunque la predicción realizada mejora la predicción de otros modelos más simples. En cualquier caso, el modelo cuenta con una amplia variedad de parámetros sujetos a posibles modificaciones para intentar optimizar el proceso de predicción, lo cual se podría llevar a cabo en futuras experimentaciones.

Así mismo, la selección de variables no es una tarea sencilla de cara a la predicción, y el propio algoritmo genético que se encarga de dicha selección puede ser modificado o sustituido para realizar un proceso de selección de variables más adecuado.

En cualquier caso, queda constancia de la complejidad que tiene realizar predicciones sobre conjuntos de datos aleatorios, ya que éstos pueden influir tanto en el entrenamiento del modelo como en los resultados de las pruebas. Las limitaciones de tiempo y recursos también juegan un papel importante en la toma de decisiones, hasta el punto que pueden llegar a limitar la eficiencia del modelo de predicción.

5

Conclusiones y Trabajo Futuro

5.1. Conclusiones

Este trabajo ha propuesto una forma de mejorar el modelo de predicción de partidos de baseball desarrollado en el trabajo anterior, Combining clustering and time series for baseball forecasting[1]. Para llevar a cabo este proceso, se ha utilizado un algoritmo genético de cara a la selección de variables que afectan a la predicción. Las variables analizadas corresponden a estadísticas de partidos de baseball de Las Grandes Ligas de Baseball, con las que se generan series temporales que sirven de base para el modelo de predicción. El modelo de predicción original ha sido adaptado para complementarse con las características del algoritmo genético.

Tras el proceso de selección de variables, se han realizado experimentaciones con los distintos conjuntos de métricas seleccionados. Los resultados de estas experimentaciones han sido analizados y se ha comparado la eficiencia del modelo de predicción propuesto con respecto a otros modelos de predicción. De estos resultados se han sacado las siguientes conclusiones:

- El nuevo modelo de predicción se ve afectado irregularmente por los conjuntos de métricas seleccionados. Esto se debe a que el uso del algoritmo genético no se traduce en una mejora sustancial del modelo, sino que se aprecia un sobreentrenamiento del modelo de predicción.
- Distintos conjuntos de métricas muestran distintos comportamientos de cara a la predicción. Algunos conjuntos de métricas muestran porcentajes de aciertos relativamente alto sobre ciertos conjuntos de datos y bajos sobre otros, mientras que otros conjuntos de métricas obtienen resultados más estables.
- El modelo de predicción aún tiene mucho margen de mejora, lo que lo convierte en una herramienta interesante de estudiar. Distintos parámetros pueden ser modificados y estudiados en futuras aproximaciones de cara a mejorar la eficiencia de la predicción.
- Aunque en este caso el uso de un algoritmo genético no ha conllevado una mejora del modelo, también pueden modificarse, no solo los parámetros del algoritmo, sino el algoritmo en sí y la forma de aplicarlo, para que se optimice la predicción.

Es importante mencionar que el tiempo necesario para realizar un estudio de este calibre, además de la importancia de los recursos tecnológicos de los que se dispone, limitan considera-

blemente la realización de ciertos experimentos, sobre todo cuando se trabaja con conjuntos de datos de un tamaño considerable.

5.2. Trabajo futuro

Al igual que para el trabajo anterior, el modelo de predicción aún tiene muchas características que son potencialmente mejorables:

- Como ya se ha mencionado, diferentes parámetros pueden ser modificados a la hora de experimentar con el modelo, y, aunque en este trabajo se ha buscado mejorar el ratio de acierto del modelo modificando algunas de ellas, aún se puede realizar mucha más experimentación en cuanto a modificación de parámetros.
- El modelo ha sido simplificado sustituyendo el clustering para realizar una experimentación viable. Con mejores recursos, y más tiempo se podría repetir este experimento con un modelo que utilice clustering y comparar los resultados obtenidos por ambos modelos.
- Experimentaciones con otros parámetros del algoritmo genético, como por ejemplo probar diferentes funciones de selección, de mutación o de cruce, pueden mejorar la búsqueda de métricas óptimas o también aplicarse para otros parámetros de la predicción para mejorar el porcentaje de acierto obtenido por el modelo.
- El modelo en sí solo tiene en cuenta estadísticas de equipos para predecir partidos, pero se podrían tener en cuenta otros factores como las condiciones del partido o las estadísticas a nivel de jugador.
- Por último adaptar y probar este modelo de predicción para otros conjuntos de datos, ya sean de baseball o de otros deportes. La dificultad de esto se debe a que existen pocos proveedores de datos deportivos que puedan dar una información tan rica como lo hace Retrosheet.

Glossary

- **MP**: Modelo de Predicción
- **DM**: Minería de Datos (Data Mining)
- **ST**: Serie Temporal
- **AG**: Algoritmo Genético
- **BB**: Baseball
- **Sim**: Similitud
- **Diss**: Disimilitud

Bibliografía

[1] Miguel Vázquez Fernández de Lezeta. Combining clustering and time series for baseball forecasting. *TFG*, 2014.

[2] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[3] Kyung-Shik Shin and Yong-Joo Lee. A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, 23(3):321 – 328, 2002.

[4] RobertP. Schumaker, OsamaK. Solieman, and Hsinchun Chen. *Sports Data Mining*. Integrated Series in Information Systems. Springer US, 2010.

[5] HéctorD. Menéndez, Miguel Vázquez, and David Camacho. Mixed clustering methods to forecast baseball trends. In David Camacho, Lars Braubach, Salvatore Venticinque, and Costin Badica, editors, *Intelligent Distributed Computing VIII*, volume 570 of *Studies in Computational Intelligence*, pages 175–184. Springer International Publishing, 2015.

[6] Gordon Waitt. Social impacts of the sydney olympics. *Annals of Tourism Research*, 30(1):194 – 215, 2003.

[7] Choong-Ki Lee and Tracy Taylor. Critical reflections on the economic impact assessment of a mega-event: the case of 2002 {FIFA} world cup. *Tourism Management*, 26(4):595 – 603, 2005.

[8] Brenda G Pitts and David Kent Stotlar. *Fundamentals of sport marketing*. Fitness information technology Morgantown, WV, 2002.

[9] Nidhal Ben Abdelkrim, Saloua El Fazaa, and Jalila El Ati. Time–motion analysis and physiological data of elite under-19-year-old basketball players during competition. *British Journal of Sports Medicine*, 41(2):69–75, 2007.

[10] Steven Pinch and Nick Henry. Paul krugman's geographical economics, industrial clustering and the british motor sport industry. *Regional Studies*, 33(9):815–827, 1999.

[11] David Forrest and Robert Simmons. Sport and gambling. *Oxford Review of Economic Policy*, 19(4):598–611, 2003.

[12] Brenda G Pitts, Lawrence W Fielding, and Lori K Miller. Industry segmentation theory and the sport industry: Developing a sport industry segment model. *Sport Marketing Quarterly*, 3(1):15–24, 1994.

[13] Yong Jae Ko and Donna L Pastore. A hierarchial model of service quality for the recreational sport industry. *Sport Marketing Quarterly*, 14(2), 2005.

[14] James M Gladden and Daniel C Funk. Developing an understanding of brand associations in team sport: Empirical evidence from consumers of professional sport. *Journal of Sport management*, 16(1), 2002.

[15] Hans H Bauer, Nicola E Stokburger-Sauer, and Stefanie Exler. Brand image and fan loyalty in professional team sport: A refined model and empirical assessment. *Journal of sport Management*, 22(2), 2008.

[16] Albert V Carron, Michelle M Colman, Jennifer Wheeler, and Diane Stevens. Cohesion and performance in sport: A meta analysis. *Journal of Sport & Exercise Psychology*, 24(2), 2002.

[17] GC Roberts and Y Ommundsen. Effect of goal orientation on achievement beliefs, cognition and strategies in team sport. *Scandinavian journal of medicine & science in sports*, 6(1):46–56, 1996.

[18] Wade D Gilbert and Pierre Trudel. Role of the coach: How model youth team sport coaches frame their roles. *Sport Psychologist*, 18(1), 2004.

[19] William McTeer, Phillip G White, Sheldon Persad, et al. Manager/coach mid-season replacement and team performance in professional team sport. *Journal of Sport Behavior*, 18(1):58–68, 1995.

[20] Joseph Baker, Jeane Cote, and Bruce Abernethy. Sport-specific practice and the development of expert decision-making in team ball sports. *Journal of applied sport psychology*, 15(1):12–25, 2003.

[21] G.M. Verrall, Y. Kalairajah, J.P. Slavotinek, and A.J. Spriggins. Assessment of player performance following return to sport after hamstring muscle strain injury. *Journal of Science and Medicine in Sport*, 9(1–2):87 – 90, 2006.

[22] Iqbal Surve, Martin P. Schwellnus, Tim Noakes, and Carl Lombard. A fivefold reduction in the incidence of recurrent ankle sprains in soccer players using the sport-stirrup orthosis. *The American Journal of Sports Medicine*, 22(5):601–606, 1994.

[23] Bradley Wilson, Constantino Stavros, and Kate Westberg. Player transgressions and the management of the sport sponsor relationship. *Public Relations Review*, 34(2):99 – 107, 2008. Special Issue: Public Relations and Sport.

[24] Hector Menendez, Gema Bello-Orgaz, and David Camacho. Extracting behavioural models from 2010 fifa world cup. *Journal of Systems Science and Complexity*, 26:43–61, 2013.

[25] Roberto N. Onody and Paulo A. de Castro. Complex network study of brazilian soccer players. *Phys. Rev. E*, 70:037103, Sep 2004.

[26] B Dawson, R Hopkinson, B Appleby, G Stewart, and C Roberts. Player movement patterns and game activities in the australian football league. *Journal of Science and Medicine in Sport*, 7(3):278 – 291, 2004.

[27] Inderpal Bhandari, Edward Colet, Jennifer Parker, Zachary Pines, Rajiv Pratap, and Krishnakumar Ramanujam. Advanced scout: Data mining and knowledge discovery in nba data. *Data Mining and Knowledge Discovery*, 1(1):121–125, 1997.

[28] E. Bittner, A. NuBbaumer, W. Janke, and M. Weigel. Self-affirmation model for football goal distributions. *EPL (Europhysics Letters)*, 78(5):58002, 2007.

[29] Pedro O.S. Vaz de Melo, Virgilio A.F. Almeida, and Antonio A.F. Loureiro. Can complex network metrics predict the behavior of nba teams? In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 695–703, New York, NY, USA, 2008. ACM.

[30] Taylor Raines, Milind Tambe, and Stacy Marsella. Automated assistants to aid humans in understanding team behaviors. In Manuela Veloso, Enrico Pagello, and Hiroaki Kitano, editors, *RoboCup-99: Robot Soccer World Cup III*, volume 1856 of *Lecture Notes in Computer Science*, pages 85–102. Springer Berlin Heidelberg, 2000.

[31] Z. Ivankovic, M. Rackovic, B. Markoski, D. Radosav, and M. Ivkovic. Analysis of basketball games using neural networks. In *Computational Intelligence and Informatics (CINTI), 2010 11th International Symposium on*, pages 251–256, Nov 2010.

[32] Biao Xu. Prediction of sports performance based on genetic algorithm and artificial neural network. *JDCTA: International Journal of Digital Content Technology and its Applications*, 6(22):141–149, 2012.

[33] Andy Cox and John Stasko. Sportsvis: Discovering meaning in sports statistics through information visualization. In *Compendium of Symposium on Information Visualization*, pages 114–115. Citeseer, 2006.

[34] Jahn K Hakes and Raymond D Sauer. An economic evaluation of the moneyball hypothesis. *The Journal of Economic Perspectives*, 20(3):173–185, 2006.

[35] Max Marchi and Jim Albert. *Analyzing Baseball Data with R.* CRC Press, Taylor and Francis Group, 2013.

[36] Xavier Golay, Spyros Kollias, Gautier Stoll, Dieter Meier, Anton Valavanis, and Peter Boesiger. A new correlation-based fuzzy logic clustering algorithm for fmri. *Magnetic Resonance in Medicine*, 40(2):249–260, 1998.

[37] Pablo Montero and José A. Vilar. Tsclust: An r package for time series clustering. *Journal of Statistical Software*, 62(1), 12 2014.

[38] Luca Scrucca. Ga: A package for genetic algorithms in r. *Journal of Statistical Software*, 53(4):1–37, 4 2013.

A

Métricas

Este anexo muestra las métricas utilizadas durante la experimentación de este trabajo, como se explica en la Sección 4. Las métricas provienen de relaciones estadísticas utilizadas por los analistas de baseball involucrados en Las Grandes Ligas de Baseball. A continuación se muestran las 40[1] métricas elegidas:

- **Runs (R & RP):** Carreras.
- **Hits (H & HP):** Número de veces en las que el bateador alcanza al menos una base tras el bateo.
- **Doubles (D & DP):** Número de veces en las que el bateador alcanza la segunda base tras el bateo.
- **Triples (T & TP):** Número de veces en las que el bateador alcanza la segunda base tras el bateo.
- **Home Runs (HR & HRP):** Jonrones.
- **Total Bases (TB & TBP):** Total de bases alcanzadas por los bateadores.
- **Runs Batted In:** Carreras obtenidas por jugada de bateo.
- **Batting Average (BA & OBA):** Porcentaje de bateo, o veces que se consigue un Hit por veces que se batea.
- **On-base percentage Plus Slugging (OPS & OPSP):** Número de veces en las que el bateador alcanza al menos una base, partido por el número de veces que batea, más las bases alcanzadas por bateo entre el número de veces que batea.
- **Strikes Out (SO & SOP):** Número de veces en las que el bateador es eliminado por errores en el bateo.

[1]Para algunas estadísticas se toma en cuenta tanto el valor propio, como el valor del equipo contrario. Por ejemplo, Runs (R) es el número de carreras propias, mientras que Runs Pitched (RP), es el número de carreras del equipo contrario.

- **Base on Balls (BB & BBP):** Número de veces en las que el bateador alcanza una base por errores de lanzamiento.
- **Stolen Bases (SB & SBA):** Bases Robadas.
- **Caught Stealing (CS & CSP):** Número de veces en las que un corredor es eliminado cuando intentaba robar una base.
- **Stolen Bases Percentage (SBP & SBPP):** Número de bases robadas entre el número de intentos de robo de base.
- **Ground Outs (GO):** Eliminaciones del bateador por jugada de pelota en tierra.
- **Ground into Double Plays (GDP):** Eliminaciones de bateador y corredor por doble jugada de pelota en tierra.
- **Fly Outs (FO):** Eliminaciones de bateador por pelota aérea.
- **Extra Base Hits (XBH):** Número total de bases alcanzadas por bateo.
- **Wins (W):** Número de veces que un lanzador es lanzador ganador.
- **Earned Runs (ER):** Carreras atribuidas al lanzador.
- **Earned Run Average (ERA):** Número de carreras atribuidas al lanzador sobre el número de innings en los que ha lanzado.
- **Walks and Hits per Inning Pitched (WHIP):** Número de veces que se permite a un bateador avanzar por errores de lanzamiento o por jugada de bateo entre el número de innings lanzados.
- **Kicks Per Walks (KPW):** Ratio de eliminaciones por lanzamiento sobre avances permitidos por errores de lanzamiento.
- **Put Outs (PO):** Eliminaciones por jugada de campo.
- **Assists (A):** Asistencias.
- **Errors (E):** Errores de campo.
- **Fielding Percentage (FPCT):** Eliminaciones y asistencias por jugada de campo.

B
Métricas

Este anexo contiene tablas con información detallada de los resultados obtenidos durante la experimentación, utilizando los distintos conjuntos de métricas se explican en la Sección 4.

Conjunto de métricas	Temporada	Media	Máximo	Mínimo	Desviación Típica
M1	2005	57.52 %	62.11 %	47.31 %	±0.046
M1	2008	52.40 %	59.00 %	48.00 %	±0.034
M1	2011	50.90 %	60.00 %	39.00 %	±0.068
M2	2005	55.23 %	58.95 %	51.06 %	±0.029
M2	2008	52.60 %	65.00 %	45.00 %	±0.069
M2	2011	52.60 %	60.00 %	45.00 %	±0.041
M3	2005	54.11 %	63.92 %	44.57 %	±0.058
M3	2008	55.90 %	63.00 %	45.00 %	±0.048
M3	2011	48.80 %	55.00 %	42.00 %	±0.044
M4	2005	53.18 %	58.33 %	48.31 %	±0.032
M4	2008	51.60 %	60.00 %	43.00 %	±0.055
M4	2011	54.10 %	61.00 %	47.00 %	±0.053
M5	2005	49.49 %	56.18 %	40.62 %	±0.05
M5	2008	51.70 %	58.00 %	42.00 %	±0.043
M5	2011	49.10 %	56.00 %	43.00 %	±0.043
M6	2005	55.11 %	61.54 %	50.00 %	±0.038
M6	2008	55.70 %	66.00 %	46.00 %	±0.055
M6	2011	52.30 %	57.00 %	46.00 %	±0.039
M7	2005	54.52 %	62.77 %	47.87 %	±0.044
M7	2008	54.50 %	65.00 %	47.00 %	±0.051
M7	2011	54.60 %	64.00 %	48.00 %	±0.049
M8	2005	52.20 %	57.89 %	48.91 %	±0.028
M8	2008	56.10 %	62.00 %	51.00 %	±0.04
M8	2011	53.00 %	62.00 %	48.00 %	±0.046
M9	2005	53.86 %	65.98 %	43.48 %	±0.067
M9	2008	50.60 %	62.00 %	42.00 %	±0.069
M9	2011	50.70 %	59.00 %	41.00 %	±0.051
M10	2005	53.34 %	59.78 %	45.65 %	±0.049
M10	2008	54.60 %	64.00 %	48.00 %	±0.063
M10	2011	50.50 %	59.00 %	41.00 %	±0.062
M11	2005	49.02 %	54.74 %	41.67 %	±0.045
M11	2008	50.30 %	58.00 %	36.00 %	±0.062
M11	2011	48.70 %	53.00 %	38.00 %	±0.043
M12	2005	56.41 %	61.70 %	48.94 %	±0.049
M12	2008	55.50 %	62.00 %	50.00 %	±0.037
M12	2011	51.00 %	59.00 %	44.00 %	±0.051

Tabla B.1: Resultados del experimento por temporada y métricas (M1-M12).

Conjunto de métricas	Temporada	Media	Máximo	Mínimo	Desviación Típica
M13	2005	58.43 %	70.45 %	48.39 %	±0.065
M13	2008	54.00 %	61.00 %	48.00 %	±0.046
M13	2011	51.00 %	63.00 %	44.00 %	±0.065
M14	2005	53.84 %	62.11 %	46.24 %	±0.056
M14	2008	57.20 %	63.00 %	50.00 %	±0.038
M14	2011	48.80 %	56.00 %	46.00 %	±0.034
M15	2005	51.46 %	62.50 %	42.39 %	±0.063
M15	2008	52.90 %	73.00 %	37.00 %	±0.098
M15	2011	50.80 %	59.00 %	42.00 %	±0.058
M16	2005	56.35 %	61.29 %	49.46 %	±0.036
M16	2008	53.10 %	58.00 %	48.00 %	±0.038
M16	2011	49.20 %	56.00 %	42.00 %	±0.044
M17	2005	52.24 %	56.04 %	43.62 %	±0.038
M17	2008	55.70 %	64.00 %	48.00 %	±0.054
M17	2011	52.40 %	61.00 %	40.00 %	±0.057
M18	2005	52.96 %	61.70 %	49.47 %	±0.045
M18	2008	56.10 %	63.00 %	49.00 %	±0.043
M18	2011	52.60 %	61.00 %	45.00 %	±0.048
M19	2005	53.72 %	61.54 %	45.83 %	±0.056
M19	2008	54.10 %	59.00 %	44.00 %	±0.046
M19	2011	52.20 %	59.00 %	45.00 %	±0.052
M20	2005	53.78 %	61.70 %	41.30 %	±0.062
M20	2008	54.00 %	64.00 %	39.00 %	±0.076
M20	2011	47.50 %	54.00 %	42.00 %	±0.039
M21	2005	55.78 %	62.50 %	51.69 %	±0.038
M21	2008	56.50 %	61.00 %	52.00 %	±0.035
M21	2011	48.50 %	57.00 %	43.00 %	±0.043
M22	2005	52.20 %	56.67 %	40.86 %	±0.047
M22	2008	55.90 %	67.00 %	50.00 %	±0.055
M22	2011	52.30 %	56.00 %	48.00 %	±0.029
M23	2005	51.96 %	61.05 %	35.48 %	±0.072
M23	2008	56.20 %	62.00 %	51.00 %	±0.035
M23	2011	49.60 %	59.00 %	40.00 %	±0.056
M24	2005	53.64 %	60.00 %	44.94 %	±0.045
M24	2008	52.40 %	57.00 %	45.00 %	±0.04
M24	2011	51.40 %	57.00 %	38.00 %	±0.055
M25	2005	54.88 %	63.83 %	43.62 %	±0.06
M25	2008	50.60 %	57.00 %	39.00 %	±0.054
M25	2011	49.90 %	54.00 %	43.00 %	±0.04
M26	2005	53.06 %	60.00 %	46.15 %	±0.049
M26	2008	54.40 %	60.00 %	49.00 %	±0.044
M26	2011	53.90 %	64.00 %	46.00 %	±0.048

Tabla B.2: Resultados del experimento por temporada y métricas (M13-M26).